

# P

## Parabolic Trough Solar Technology

ECKHARD LÜPFERT

DLR Deutsches Zentrum für Luft- und Raumfahrt  
German Aerospace Center, Institute of Technical  
Thermodynamics, Solarforschung, Koeln/Cologne,  
Germany

### Article Outline

Glossary  
Definition of the Subject and Its Importance  
Introduction  
Technology Description  
Future Directions  
Bibliography

### Glossary

**Collector loop (PTC)** Unit of several PTC connected in series to heat the fluid from inlet to outlet header temperature.

**Direct normal irradiance, beam irradiance** Direct part of the sunlight, coming from within the sun disk as almost parallel light onto a surface, measured as power density in  $\text{kW/m}^2$ .

**Drive (PTC)** Unit consisting of motor and gear or hydraulic drive with valves and cylinders, and the controller to turn the PTC into the correct operational tracking angle.

**Efficiency** Ratio of useful energy and total energy input.

**Efficiency (PTC)** Ratio of thermal energy output from the PTC and total solar radiation received on the aperture area.

**Heat transfer fluid (“HTF”)** Fluid receiving the thermal energy in the receivers and transporting it to the heat exchangers, etc., of the power block. HTF for PTC is mostly synthetic oil or water/steam.

Some installations also use molten salt or pressurized  $\text{CO}_2$ .

**Intercept factor** Relative amount of rays hitting the absorber tube as fraction of the total number of reflected rays from the mirror area.

**Mirror, mirror panel (PTC)** Reflecting panel, made of silvered glass or other reflecting sheet material, curved to reflect sunlight onto the absorber.

**Module (PTC)** Parabolic trough collector section between pylons, including structure, mirrors, and receivers.

**Parabolic trough collector (“PTC”)** Concentrating solar collector with mirrors, absorber, and tracking system for providing solar energy at temperatures of  $100\text{--}600^\circ\text{C}$ .

**Pylon (PTC)** Support post of the PTC modules.

**Receiver (PTC)** Component of a concentrating collector system, especially PTC, consisting of absorber tube, with additional elements such as glass tube and expansion bellow.

**Solar field** Unit of parallel connected collector loops, typically also including connection piping, sensors and controls, land area, and heat transfer fluid of the collector installation of a solar plant.

**Sun sensor** Sensor for feedback of the tracking to the drive.

**Tracking (PTC)** Action of adjusting the collector angle to the sun position during the operation.

### Definition of the Subject and Its Importance

Parabolic trough (solar) collectors (PTCs) are technical devices to collect the energy in form of solar radiation and convert it typically into thermal energy at temperature ranges of  $150\text{--}500^\circ\text{C}$  at industrial scale. The cylindrical trough shape of the reflecting surface with parabolic section of the mirror shape has the ability to concentrate the incident sunlight onto an absorber tube

in the focal line of the collectors. Typical width of such PTC is 0.5–10 m. Main use of PTC is in solar power generation. In large-scale concentrating solar power applications, the PTC is the most successful type of concentrating collector design. The first troughs are reported at the end of the nineteenth and beginning of the twentieth century for industrial-scale steam generation. The wide expansion of coal, oil, and gas for heat and power generation left solar energy technology behind until oil price shocks initiated a development step in the 1980s, leading to the successful commercial start of the parabolic trough solar power plants SEGS I–IX in California until 1990. Larger scale capacities have been installed in Spain since 2007, and from there and since then spreading out worldwide.

PTCs are the main technology for large-scale concentrating solar power (CSP) solar fields. The increasing application in CSP is due to the high conversion efficiency in combination with a standardized modular design with relevant economic advantages over other CSP variants.

## Introduction

PTCs feature a concentrator shell with parabolic sections in a cylindrical configuration. The focus of the cylinder parabola is a straight line. The aperture width is typically about three to four times longer than the focal distance of the parabola between vertex and absorber tube (Figs. 1 and 8).

A PTC consists of modules; these are the units supported between pylons.

Main components of a PTC module are the curved mirror and the absorber tube. Its support structure holds the components in place and connects the modules between each other.

PTCs are among the most efficient available concentrating solar collectors, only superseded by the paraboloidal dish concentrators (reference to Article). The advantage of the PTC increases in larger installations, where the collection of the heat from a solar field with many collectors to a central conversion system is needed with a wide system of heat transfer pipes.

Parabolic trough collectors are tracking reflector systems. There is no reasonable chance to avoid the

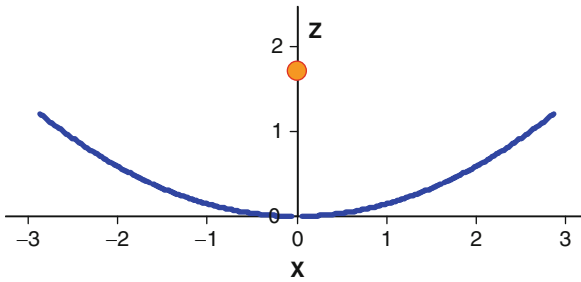
continuous rotation of the collector to always point with the optical axis (symmetry axis) plane to the sun. The tracking accuracy must be typically in the range of  $0.1^\circ$ . This order of magnitude applies to all orientations of any of the components of the concentrating collector.

The sun is the energy source for PTC. The sun apparently moves daily over the sky, due to a rotation speed of the earth axis of  $15^\circ/\text{h}$ . The sun changes its path over the sky throughout the year due to the tilt of the earth's rotation axis of  $\pm 23^\circ$  from the rotation plane around the sun. Nevertheless, PTCs require only one horizontal rotation axis for the tracking. The tracking axis is typically oriented north–south for commercial application, as this results in the best output over the year in the typical latitudes of PTC application.

Typical dimensions of a PTC are given in Table 1 and illustrated in Fig. 1.

**Parabolic Trough Solar Technology. Table 1** PTC dimensions of a typical CSP plant, e.g., EuroTrough

Collector width	5.78 m
Collector length	150 m
Focal distance	1.71 m
Module length	12.3 m
Number of modules per collector	12
Effective aperture per collector	818 m <sup>2</sup>
Absorber tube diameter	70 mm
Ideal concentration factor	82×
Geometric concentration factor	26×
Absorber tube length	4.06 m
Glass tube diameter	120 mm
Number of panels per section	4
Mirror panel size	1.64 × 1.7 m <sup>2</sup>
Mirror thickness	3.8 mm
Solar weighted reflectance of mirror	>94%
Curvature accuracy of collector module	<3 mrad
Specific weight without foundations	<30 kg/m <sup>2</sup>
Operational wind speed	<15 m/s
Max wind speed (survival position), typical	38 m/s



**Parabolic Trough Solar Technology. Figure 1**  
PTC section – dimensions of EuroTrough, LS-3, and other designs, in meters

### Technology Description

Main components of a parabolic trough collector module are the curved mirror and the absorber tube. Its support structure holds the components in place and connects the modules between each other.

#### PTC Mirror

Mirrors are required to directionally reflect the incoming sunlight onto the absorber tube. Relevant properties of the mirrors are thus:

1. High direct (specular) reflectance (evaluated for the solar spectrum, e.g., ASTM G-173 direct+circumsolar). Typical benchmark values are >92% of solar-weighted direct reflectance.
2. Shape fidelity of the parabolic shape. Deviations produce an inaccurate focus and can lead to “spill-age.” Goal is an intercept factor of >95% over the majority of the annual operating conditions.
3. Shape stiffness under dead load and wind load. Deformations are allowed for high wind speed, as they occur only in a limited number of operating hours of the collector.
4. Durability in terms of surface quality (hard coat, not scratched) and in terms of reflector quality. The preferred reflecting material is silver due to its good optical properties for reflecting the solar spectrum. Aluminum and multilayer coatings are under development. Metallic coatings are sensitive to corrosion and require appropriate long-life protection.

Typical CSP mirrors are made from float glass of 3–4 mm thickness with low iron content to avoid absorption in the glass, with a silvered backside,

protected by a multilayer protection coating from the backside (copper, protection paints, and final lacquer). Options from aluminum and polymers typically suffer from lower optical, mechanical, and durability properties.

#### Absorber (Receiver)

The absorber of a PTC usually is a dark-coated metal tube to absorb the incoming solar radiation. It absorbs the concentrated radiation reflected from the mirror, and also the global radiation hitting from top. The absorber tube optical properties are preferably selective, with absorbance of 95% and more in the solar spectrum range (300–2,500 nm) and low emittance for the infrared radiation (beyond 2  $\mu\text{m}$ ) to reduce the thermal losses in operation. This is achieved with sputtered Cermet coatings consisting of several layers of metallic and ceramic coatings. Also, galvanic black-nickel and black-chrome coatings are applied but have less temperature resistance and higher emissivity.

The absorber tube must be protected by a glass tube to reduce convective heat losses and to protect the sensitive absorber surface from soiling and mechanical damage. For temperatures above 250°C, it is useful to have the gap between absorber and glass tube evacuated. This reduces oxidation degradation of the surface and eliminates convective heat losses. However, such design requires sophisticated geometries to compensate the thermal expansion of the absorber tube and keep a long-life vacuum stable. Critical elements are the bellow and joint between glass and metal tubes, to avoid thermal tensions that lead to glass damage. Careful design and extensive tests are required. Additional stabilizer for the vacuum conditions is a so-called getter from reactive metal such as barium. An indicator color change to white identifies vacuum loss.

Most widespread geometry is the receiver, originally developed by Luz Industries, with 70-mm absorber tube diameter and 4,060-mm length at ambient temperature (Table 2).

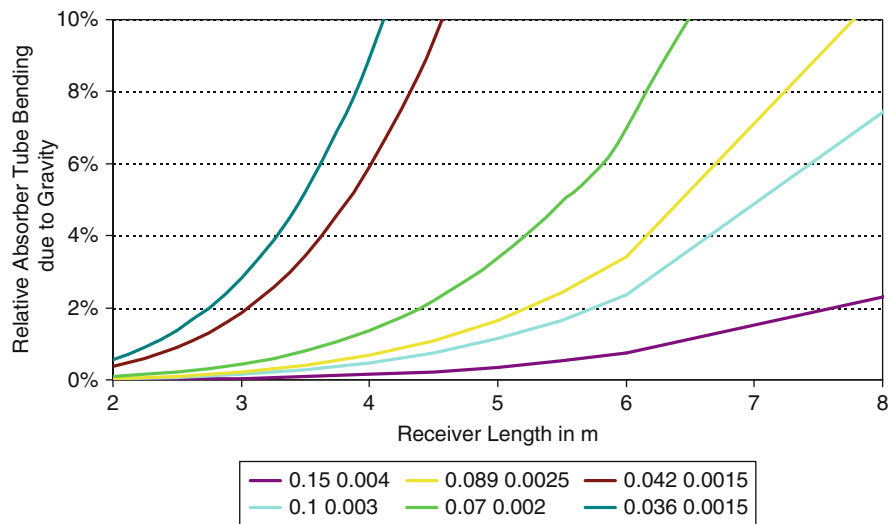
The length of the receiver tube between supports is limited by the bending of the absorber tube under dead weight of tube and contained fluid (Fig. 2). A bending of about 2–5% of the absorber tube diameter can be accepted. More bending would lead to losses in intercept factor. Resulting distance between supports

**Parabolic Trough Solar Technology. Table 2** PTC receiver typical properties (UVAC, PTR70, HEMS, and others)

	Typical	Range
Absorber tube diameter	70 mm	10–110 mm
Absorber tube wall thickness	2 mm (2–6 mm)	2 mm (1–10 mm)
Receiver length (cold/hot)	4.06/4.08 m	2–6 m
Glass tube outer diameter	115–125 mm	
Effective useful absorber length (hot)	96.5%	94–97%
Glass transmittance	96%	>90%
Absorptance	96%	>94%
Overall optical efficiency	88%	80–90%
Heat loss at 350°C	180 W/m	
	0.9 kW/m <sup>2</sup>	
Insulation	Vacuum	
Expected lifetime	>25 years	

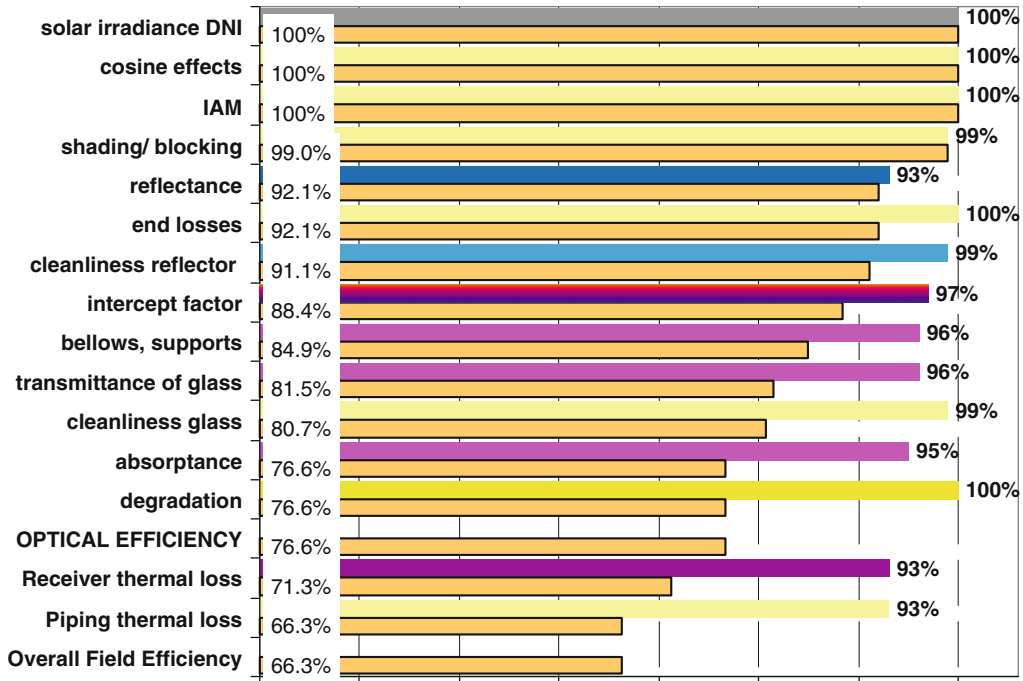
strongly depends on the area inertia of the tube, in particular, on the tube diameter, and is typically of 3–5 m maximum.

**Performance Properties of Receivers for PTC** The optical performance of receivers is determined by the properties of the absorber coating and of the surrounding glass tube. Absorptance of the absorber coating is determined by measuring spectral reflectance and calculating the solar spectrum weighted value. Typical range at current state of the art is 95–96%. New coatings are under development to further increase the properties. However, the requirement of low emittance in the infrared range sets limits in the range of overlap between solar spectrum (gains) and thermal radiation (loss) in the range of 1.5–2.5  $\mu\text{m}$ . Very low reflectance is desired at shorter wavelengths (sunlight), and very high reflectance at longer wavelengths (infrared). These properties are, however, difficult to measure on an intact receiver. For measuring the performance characteristics of such receivers, DLR has developed laboratory tests for receivers for the heat loss properties and their optical efficiency [14, 15] and applies them in



**Parabolic Trough Solar Technology. Figure 2**

Absorber tube bending relative to absorber diameter for stainless steel tubes for 40 bars, welded ends, filled with oil, at operating temperature, for receiver geometry layout



Parabolic Trough Solar Technology. Figure 3  
Efficiency diagram for high performance trough collector field



Parabolic Trough Solar Technology. Figure 4  
Parabolic trough collector (EuroTrough) solar field installation



**Parabolic Trough Solar Technology. Figure 5**  
Parabolic trough collector (EuroTrough) solar field installation



**Parabolic Trough Solar Technology. Figure 6**  
Andasol-1 solar power plant in Granada (Spain)

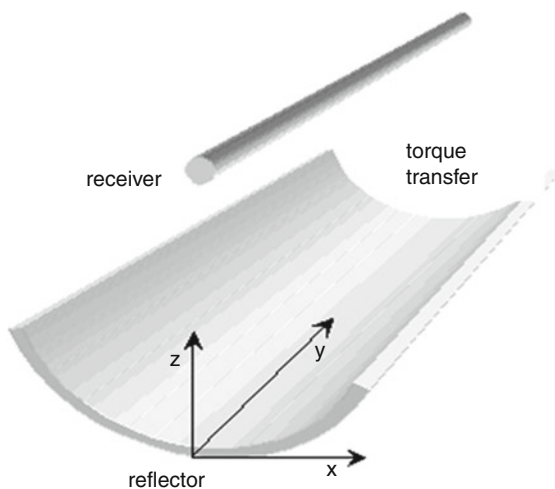
its QUARZ Center, with other test labs preparing similar tests and test standards.

Optical efficiency is the product of area factor  $\times$  transmittance  $\times$  absorptance. It may depend on incidence angle and absorber temperature. Heat loss

depends on absorber temperature and is tested for constant homogeneous absorber temperature. The test result is a heat loss curve over absorber temperature and specific values in  $\text{W/m}$  or  $\text{kW/m}^2$  for standard temperatures of  $350^\circ\text{C}$ ,  $400^\circ\text{C}$ , etc.



**Parabolic Trough Solar Technology. Figure 7**  
Parabolic trough solar field, La Florida (Extremadura), Spain



**Parabolic Trough Solar Technology. Figure 8**  
Main elements of parabolic trough collector

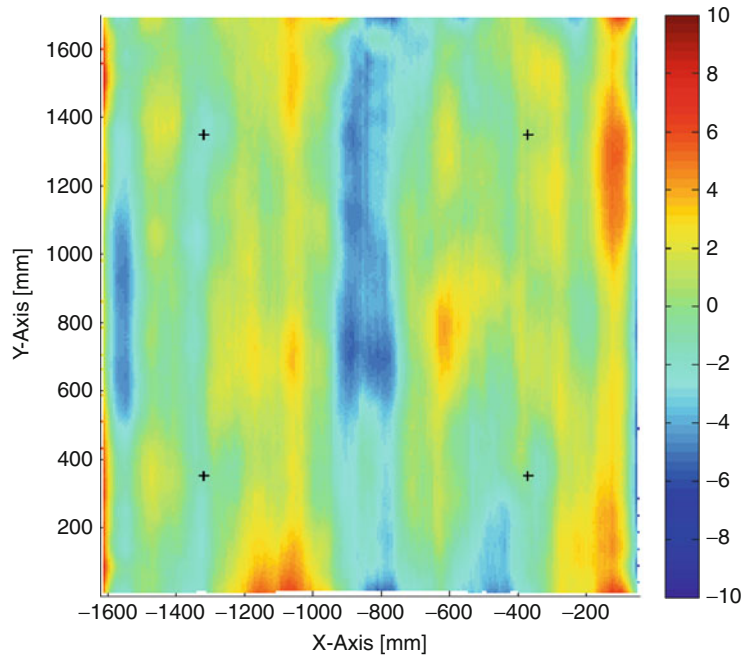
### Support Structure

The PTC support structure carries the key components of the collector module and is responsible for their accurate alignment in any operating conditions and over an operation period of several decades. Main criteria beside geometric accuracy and cost effectiveness are thus mechanical stiffness and corrosion resistance. The loads affecting the structure include dead load, weight of the components, loads of the receivers, and wind loads (bending and torque).

The structure needs to be designed in a modular way to ease the assembly of the collectors and proper alignment in the solar field.

### Tracking<sup>3</sup> and Drive Mechanism

The concentrating properties of the collector require tracking of the PTC to the apparent movement of the sun over the sky. The PTC as on-axis concentrator



**Parabolic Trough Solar Technology. Figure 9**  
Mirror shape measurement result from QDec system, slope deviation  $SD_x$  in mrad



**Parabolic Trough Solar Technology. Figure 10**  
Parabolic trough test platform at Plataforma Solar de Almería, Spain





**Parabolic Trough Solar Technology. Figure 11**  
Parabolic trough test platform at PSA, Almería, Spain



**Parabolic Trough Solar Technology. Figure 12**  
EuroTrough support structure



**Parabolic Trough Solar Technology. Figure 13**

Solar field layout configuration, as seen from space (SEGS XIII and SEGS IX, California, Source: GoogleEarth)

geometry requires that the sun is on the plane spanned by the optical axis and absorber axis. The angle between normal on the aperture area of the tracked collector and the sun beam is called incident angle. Good accuracy of the tracking is required to obtain high intercept factor. The required accuracy is about 1 mrad ( $0.02^\circ$ ). The tracking controller is usually connected to a central control system linking the field control and the power block control.

The most widely used drive mechanism for PTC is based on hydraulic systems. They provide high force, slow speed, no backlash, and long-term reliability under outdoor conditions in a solar field.

The angular encoder accuracy must be below  $0.05^\circ$ . Also, differential solar sensors are used to enhance the accuracy of the tracking for reaching the tracking accuracy.

### Foundation and Pylons

Foundations, usually concrete blocks or piles, are fixing the collectors in the ground. The concrete base carries the pylons with the bearings on which the collector turns with the tracking movement. Preferred orientation of the collectors is with the horizontal rotation axis in north–south direction. Although it is common to level the field, it is not required to have a leveled

collector axis. Slopes of up to 3% toward north or south have been demonstrated with the EuroTrough collector and other similar variants. Loads to be taken into account are dead load and wind load. The foundation and pylons need to be designed for local wind load standards. Typical assumptions for wind load are for the range of 34–41 m/s as maximum design wind speed. The maximum load case should be assumed for a 30-year or 50-year wind gust, according to the local codes that are also used in other construction works in the region. Wind-tunnel testing is recommended to get additional information for optimization of the design.

The effect of wind on the collector field can be reduced drastically by turning the collectors in stow toward horizon or slightly below, to prevent the wind from passing below the collectors as well as by installing a fence or hedge or other wind protection device around the collector field, in particular, toward the main wind direction.

### Heat Transfer Fluid

Proven heat transfer fluids (HTF) for parabolic trough collectors are synthetic oil and water/steam. Typical HTF is a biphenyl/diphenyl oxide fluid with temperature stability up to  $400^\circ\text{C}$ . Its use at elevated temperature requires elevated pressure in the solar field and

heat transfer system. Lifetime of the HTF should not be limited, but recovery techniques are usually required to maintain fluid properties constant. The HTF transports the thermal energy from the solar field to the heat-exchanger steam generator systems providing the superheated steam for the turbine of typically 370–380°C. The flow rates through the receivers must be high enough to ensure appropriate heat removal from the absorber walls and low enough to keep pumping power for the fluid reasonably low. The loop length of a PTC field depends on heat transfer fluid properties and preheater outlet temperatures.

Lower temperatures up to 320°C allow the use of mineral oils with lower vapor pressure.

Water as heat transport fluid in the receivers can be used to produce steam at temperatures above 400°C. Extended pilot scale testing at Plataforma Solar de Almeria has shown the feasibility of the evaporation and superheating of steam in PTC loops successfully. First commercial installations are also using water and steam for the heat transport from the PTC field. They demonstrate that the control aspects of the direct solar steam generation are solvable at large scale, resulting in cost savings and better environmental evaluation of the technology.

### Future Directions

New developments, size, materials, fluids, field design, and standards.

### Bibliography

1. [http://www.solarpaces.org/Tasks/Task3/reflectance\\_guideline.htm](http://www.solarpaces.org/Tasks/Task3/reflectance_guideline.htm)
2. [www.helioscsp.com](http://www.helioscsp.com) and [www.protermosolar.com](http://www.protermosolar.com)
3. [www.dlr.de/sf](http://www.dlr.de/sf) and [www.dlr.de/energie](http://www.dlr.de/energie) (news, downloads, ...)
4. [www.dlr.de/tt](http://www.dlr.de/tt) (Trans-CSP and Med-CSP reports)
5. [www.estelasolar.eu](http://www.estelasolar.eu) (European association)
6. [www.solarpaces.org](http://www.solarpaces.org) (International Energy Agency, CSP implementing agr)
7. [www.nrel.gov/csp/troughnet](http://www.nrel.gov/csp/troughnet) (excellent technical info base)
8. [www.psa.es](http://www.psa.es) (Plataforma Solar de Almería)
9. [www.desertec.org](http://www.desertec.org) (news, link to 10'000 Solar Gigawatts)
10. [ec.europa.eu/research/energy/nn/nn\\_rt/nn\\_rt\\_cs/article\\_1114\\_en.htm](http://ec.europa.eu/research/energy/nn/nn_rt/nn_rt_cs/article_1114_en.htm)
11. [www.powerfromthesun.net](http://www.powerfromthesun.net) (B. Stine, Solar Technology)
12. [www.stadtwerke-juelich.de](http://www.stadtwerke-juelich.de) (1.5 MW solar tower in Germany)
13. [www.solarmillennium.de](http://www.solarmillennium.de) (developer, technical info, news)
14. Lüpfer E, Riffelmann K-J, Price H, Moss T (2006) Experimental analysis of overall thermal properties of parabolic trough receivers. SolarPaces Sevilla (Poster, A2-P4 Parabolic-Trough Collector Technology (2), ISBN 84-7834-519-1, and J Sol Energy Eng 130, 2008)
15. Pernpeintner J, Schiricke B, Lüpfer E, Lichtenthäler N, Macke A, Wiesemeyer K (2009) Combined measurement of thermal and optical properties of receivers for parabolic trough collectors: In: Solarpaces conference, Berlin

## Passive House (Passivhaus)

KATRIN KLINGENBERG

Passive House Institute US, Urbana, IL, USA

### Article Outline

Glossary

Definition of Subject

Introduction

The Early Developments in North America Inspired Europe

Recent Developments in the USA and Canada

Climate-Changing Opportunity

Principles of Passive House Design

Future Directions of Passive House in the USA, Canada, and Internationally

For More Information

### Glossary

**Airtight construction** A method of construction for building envelopes that is aimed at the lowest possible result close to zero infiltration when tested for air leakage.

**Balanced mechanical ventilation with heat recovery**

Highly energy efficient mechanical equipment for airtight homes to provide equal amounts of fresh air at all times to the inhabitants as well as to exhaust the same amount of air to remove moisture, stale air, and indoor pollutants while recovering the energy at the highest possible level, which is useful to maintain thermal comfort in the building.

**Passive house building** A building achieving close to thermal stasis by optimizing heat loss and gain through the building shell (primarily by limiting transmission losses through climate-specific insulation levels of its components and appropriate

window specifications) so that thermal comfort in winter and summer can be maintained mostly without energy input and during the peak temperature periods with only very small amounts of energy input of roughly 1 W/sqft.

**Superinsulation** Adequate, climate-specific amount of insulation to balance heat gains and losses in a cost-effective manner.

**Thermal Bridge free** A building component, for example, a stud in a wall, is considered to be thermal bridge free when it has a lower conductivity than 0.006 BTU/(ft<sup>2</sup>F).

### Definition of Subject

Performance based energy metrics and Passive Design Standards, history, development, methods and tools, current state of implementation, and future outlook.

### Introduction

The Passive House Building Energy Standard is the most ambitious energy standard in the world. Buildings account for 40–50% of the total US carbon emissions – depending on how the sector is defined – stemming from buildings. The Passive House design and construction approach proposes to slash space conditioning energy consumption of buildings by up to an amazing 90–95% and overall energy use of space conditioning and household electricity by 70–80%, depending on which reference baseline home or energy code is used for comparison. Thousands of homes as well as educational and commercial buildings have been built or remodeled to meet the Passive House standard in Europe. Widespread application of Passive House design principles and construction methods in the USA – to both new construction and retrofit scenarios – would dramatically reduce the country's energy use and reduce carbon and other harmful admissions. And, structures built to the Passive House Building Energy Standard have a high quality, long lasting envelope, and in practice provide superior comfort and indoor air quality.

The Passive House Building Energy Standard actually evolved from the superinsulated buildings of the 1970s, many of which were built and pioneered in North America. The Small Homes Council of the University of Illinois (now known as the Building Research

Council), developed the Low-Cal house in the late 1970s, and Wayne Schick, a member of the architectural faculty then, is credited with coining the term “superinsulation.” A second group of early designs incorporated passive solar heating strategies instead of superinsulation primarily.

The Canadian National Research Council and Minnesota's Housing Finance Agency were instrumental in supporting the construction of very low-energy prototypes and in launching of low energy programs like R2000. The history and list of early adopters and realized projects in Canada as well as in the USA is extensive as also the accompanying research, and would warrant a book in itself. By the end of 1986, an estimated 10,000 very low-energy homes had been built in North America. Many of these early prototypes are still around today and provide excellent long-term experience with this construction type. They are, to this day, comfortable, energy-efficient homes with very good indoor air quality (IAQ) and long-lasting, high-quality envelopes.

However, it was in Europe – in the early 1990s – that pioneering concepts like superinsulation and passive solar approaches were further refined into a comprehensive approach called Passivhaus in Germany. It was a quantum leap triggered by the passage of an ambitious, rigorous energy standard in 1988 for new buildings in Sweden. Other European countries later followed suit in tightening their energy codes significantly. In response to the new Swedish standard, Swedish professor Bo Adamson and German physicist Wolfgang Feist envisioned a building that could meet and even exceed this standard – the Passive House. Construction on the first prototype, a four-unit row house structure, started in 1989 in Darmstadt, Germany. It was completed in 1991. Then as now, the primary components of a Passive House were thick insulation, few or no thermal bridges, an airtight envelope, excellent insulated glazing, and balanced heat recovery ventilation.

### The Early Developments in North America Inspired Europe

The term “Passive House” seems to have originated in Canada and the USA. In the 1970s and 1980s, the term “Passive Housing” was a well-known scientific term

describing a combined approach of superinsulation and measured passive solar strategies minimizing space conditioning to extremely low loads. The term was translated into *Passivhaus* in German and was kept as the design principles were optimized by Feist and Adamson. They understood the term to be appropriate to describe a holistic design strategy. Tailored to the moderately cold, heating-dominated Central European climate, it aimed at requiring so little energy to heat to comfortable conditions that a conventional heating system could be eliminated. Instead, a single 1,000-W post-heating element in a balanced supply ventilation stream provided all the heating requirements for a home of approximately 1,000 ft<sup>2</sup> or 100 m<sup>2</sup>. The annual total heating demand could be minimized by making careful use of existing internal heat sources – people, lights, and appliances; and optimized (not maximized) solar energy admitted by the windows. A fresh-air heating supply would be pre-warmed by highly efficient heat recovery from the exhaust air, and also in the earlier Passive House projects in Europe by an earth tube, which is a passive geothermal heating-and-cooling system (also intended to prevent the heat exchange core of the ventilation system from freezing).

Early prototypes were designed so that the maximum heat load in the German winter was to be less than 10 Watt per square meter (W/m<sup>2</sup>), or 0.9 watts per square foot (W/ft<sup>2</sup>), of floor area. Under these circumstances, the heat load could be comfortably supplied using fresh-air ventilation – eliminating the need for a separate means of heat distribution. On the active solar and renewable side, cost-optimized solar thermal systems for hot water production were specified but there was no necessity or emphasis on any active-solar contribution such as photovoltaic. It was recognized early on that the additional investment cost for Passive House measures to conserve energy were far less and more cost-effective than active photovoltaic systems to reduce energy consumption. Passive House as a conservation baseline has since been recognized in the European Union to be the most cost-effective way to achieve near zero or zero energy buildings. Supported by research grants from the German state of Hessen, Dr. Feist created detailed computerized simulations modeling the energy behaviors of wall and window assemblies and other construction elements.

Then he systematically varied these elements to arrive at the best possible construction packages, based on energy efficiency, installation expense, and sustainability.

In 1995, Amory Lovins – himself an American energy pioneer and founder of the Rocky Mountain Institute – visited the optimized Passive House prototype at Darmstadt and was deeply impressed. Lovins had implemented Passive House principles in his own home, and saw in Feist's work more than a concept. He encouraged Feist to use the project as a basis for a practical way to meet energy needs for a broad range of implementation. All that was needed was to redesign some details to reduce construction costs.

In 1996, Dr. Feist founded the Passivhaus Institut (PHI) in Darmstadt. The PHI has flourished under his leadership, designing, testing, calculating, certifying, and analyzing data on buildings constructed to the Passive House standard, and their components. The Passive House Planning Package (PHPP) is the Institut's energy-modeling program. The methodology used in this program is profoundly thorough and balanced for interdependencies, and it has been used to design thousands of projects of all building types across Europe. The measured energy performance of the homes has been shown to closely match the modeled performance predicted. The PHPP modeling tool is the most accurate and simple-to-use design software available today to predict the energy consumption of extremely low load homes. Its granularity has been calibrated to see the smallest design effects on the energy performance of a building. Some of those effects are precise calculation of transmission losses through all components of the thermal envelope, solar gains, shading, thermal bridging effects, the influence of thermal mass, ventilation losses, systems efficiencies, and ventilation system design for minimized fresh air conditioning.

The Passive House Building Energy Standard defines the lowest energy metric and standard worldwide. It requires that a building use no more than 15 kilowatt-hours per square meter (kWh/m<sup>2</sup>) per year in heating and cooling energy, which is equivalent to 1.35 kWh/ft<sup>2</sup> or 4.75 thousand British thermal units per square foot (kBtu/ft<sup>2</sup>) annually. It further requires that the building's total primary energy consumption – that is, source energy used for space

conditioning, hot water, and electricity – not exceed 120 kWh/m<sup>2</sup> (10.8 kWh/ft<sup>2</sup> or 38 kBtu/ft<sup>2</sup>) per year.

The energy metric is in reference to the interior treated floor area minus circulation spaces and discounted storage and mechanical spaces. Passive Houses also tend to have thicker exterior envelopes than typical construction, and the energy calculation is more accurate if they are excluded. It is important to note that most modeling tools in the USA use the exterior dimensions of a building to determine the per-square-foot energy metric of a building. This has led to some confusion in comparing the Passive House energy metrics to commonly used energy intensity units in the USA. Typically, if put in relation to the exterior building dimensions – as is customary in US energy modeling, the Passive House metric is approximately 20–30% lower than the certification requirements mentioned above, depending on size and design of the building. Using exterior building dimensions, the Passive House metric would result in roughly 3.4 kBtu/ft<sup>2</sup> annually for the total space heating and cooling demand and in roughly 28 kBtu/ft<sup>2</sup> annually for the source energy requirement.

Structures built to the Passive House Building Energy Standard – and the techniques and products developed for them – were further popularized in Europe through the European Union-sponsored Cost Efficient Passive Houses as European Standards (CEPHEUS) project, which validated that the Passive House concept worked in five European countries over the winter of 2000–2001. Thousands of Passive Houses have been built across Europe, as interest in the benefits that they provide has skyrocketed. Many provinces and cities there are now mandating that all new construction built with public monies be built to the Passive House Building Energy Standard. The European Union has proposed to adopt the Passive House Building Energy Standard Europe-wide to meet their action plan, which calls for near zero energy buildings by 2020 for all new construction.

### Recent Developments in the USA and Canada

In the spring of 2002, German-born architect Katrin Klingenberg traveled from the USA, where she has lived since 1994, to Germany to tour Passive Houses with Manfred Brausem, a leading architect and Passive

House pioneer there. An ardent advocate of sustainable architecture, Klingenberg was powerfully affected by what she saw. She returned to the USA and began designing her own Passive House, the Smith House, which broke ground that October in Urbana, Illinois. In 2003, Klingenberg attended the Seventh International Passive House Conference, in Hamburg, where she met Dr. Feist. She returned to finish the Smith House, which became the first Passive House in North America that used the Passive House Planning Package software as a design tool. Monitoring devices installed at the Smith House after its completion confirmed the predictions:

- The house uses only 11 kWh/m<sup>2</sup> (1 kWh/ft<sup>2</sup> or 3.5 kBtu/ft<sup>2</sup>) per year in heating energy (in reference to the Passive House interior treated floor area).
- The highest monthly energy use ever for space and water heating, appliances, and lighting – in short, for all purposes – was 599 kWh.
- With an average electrical base load of 265 kWh, the highest monthly energy use for space heating has been 334 kWh.

The construction of the Smith House was successful in more than just one way. It confirmed the applicability and accuracy of the European-developed design tool for this North American location, as well as cost-effectiveness and relative ease of transfer to local construction techniques.

In 2005, Stephan Tanner, a Swiss-born architect started working on the first American school Passive House building, the BioHaus in Bemidji, Minnesota – located in an even more challenging climate. Designed as a learning facility for the German-language Concordia Village, the BioHaus succeeded in meeting the Passive House Building Energy Standard and was certified in 2006. In October 2006, Klingenberg and Tanner teamed up to organize the first North American Passive House conference at the BioHaus in Minnesota. Interest was piqued; new projects were started.

### Climate-Changing Opportunity

In April 2007, Klingenberg and Passive House builder Mike Kernagis cofounded the Passive House Institute United States (PHIUS) to disseminate information

about, and promote the construction of, Passive Houses and Buildings in North America. They steadily promoted the Passive House standard at symposia, workshops, and conferences from coast to coast. Passive House projects have been built in the full range of North American climates since, from Fairbanks, Alaska to Lafayette Louisiana. Although most Passive House construction to date consists of new buildings, a few retrofit projects have been started as well. By mid-2011, there were a total of 22 certified Passive House projects in the USA, with two in Canada registered. More than 100 others from the USA and Canada have applied for pre-certification and are in progress. More than 500 professionals have taken the certification training to become a Certified Passive House Consultant; the number who have taken and passed the final rigorous exam has surpassed 200.

Passive House design and construction methods and the overall approach to homebuilding best meet today's energy and environmental needs worldwide. The Passive House concept is a timely and powerful solution that is now quickly gaining traction in the USA and Canada. Many Passive House projects in this country are so new that empirical data on their energy performance are limited, but extensive data on the energy performance of Passive Houses in Europe are available, through the CEPHEUS project at [www.cephus.de/eng/index.html](http://www.cephus.de/eng/index.html). According to Guenter Lang, the former executive director of the IGPassivhaus Austria, a member-based interest group, Austria would reach a projected 25% market penetration of Passive House Buildings for new construction by the end of 2010. In the USA, various Building America teams are involved in monitoring the early Passive House projects and many have completed their first year of data collection. Soon there will be publications available on measured Passive House performances in the USA and Canada.

Early indications from that data suggest strongly that adoption of the Passive House Building Energy Standard will continue to accelerate in North America. The PHIUS, the existing North American Passive House community – active consultants organized as the National Passive House Alliance a membership organization – and future Passive House adopters together face a terrific opportunity and an enormous challenge. For the Passive House Building Energy

Standard to become the prevalent energy performance standard, it is critical that:

- Training and continuing education opportunities grow to meet the need.
- Experiences in North America's diverse climate regions are measured and shared with the building community. That experience must drive continual refinement of Passive House design principles, construction techniques, component design, and manufacture to meet the special requirements of each climate zone.
- To maintain integrity in the marketplace, rigorous project and professional certification programs must be maintained and subscribed to.

### Principles of Passive House Design

The Passive House concept is a comprehensive approach to cost-effective, high quality, healthy, and sustainable construction. It seeks to achieve two goals: minimizing energy losses and maximizing passive energy gains. A Passive House requires *up to* 95% less energy for space heating and cooling than a conventionally constructed house. To attain such outstanding energy savings, Passive House consultants and builders work together to systematically implement the following seven principles:

- Superinsulation (depending on climate).
- Eliminating thermal bridges.
- Airtight construction.
- Heat or energy recovery ventilation (depending on climate).
- High-performance windows and doors (depending on climate).
- Optimization of passive-solar and internal heat gains (depending on climate).
- Calculating the energy balance.

### Superinsulation

The insulation applied to a house works in much the same way as the insulation in a thermos bottle. In both cases, the insulating outer shell or envelope blocks or slows heat transmission and maintains the contents at a relatively constant temperature. Warm contents stay warm, cool contents stay cool, even when the

temperature on the outside hits one extreme or another. In a building constructed to the Passive House standard, the entire envelope of the building – walls, roof, and floor or basement – is well insulated.

How well insulated? That depends, of course, on the climate. To achieve the Passive House standard, a home in Sonoma, California, required only 6 in. of blown-in cellulose insulation to meet the standard, while a home in the far colder climate of Urbana, Illinois, needed 16 in. – almost three times as much. Often, the first feature of a Passive House to catch a visitor's attention is the unusual thickness of the walls. This thickness is needed to accommodate the required level of insulation.

**The Comfort Principle** Thermal comfort in summer and winter is the focus for determining the required thickness of the insulation based on climate. Human thermal comfort in a building depends on many factors, but most commonly and directly perceived according to air temperature in the space and air movement. If the mean radiant surface temperature of a building's exterior wall components is way below the surface temperature of the interior walls, then there is convection induced. If there are convective currents from warm to cold then the consequence is stratification: different temperatures near the floor of a room compared to the ceiling, or significant differences in first floor temperatures and second floor temperatures. In addition, if exterior walls are cold, then the human body starts to lose heat to the colder exterior surfaces by radiation, which leads to discomfort and feeling cold in winter. In summer, the hot exterior surfaces radiate to the interior.

There is an easy solution to this problem: maintaining the exterior surface temperature uniformly at a level so that convection is nearly eliminated. By determining the set points for the interior room temperature for summer and winter comfort, and looking at the summer and winter design temperatures in any given climate, the R-value can be calculated so that the exterior wall temperature even during the coldest design day in winter or warmest in summer will remain within the acceptable range. In a Passive House, the difference in temperature of exterior and internal wall surfaces should not exceed 4°F to always maintain human thermal comfort.

**Insulation Materials Choices** Even with this insulation requirement, Passive House designers have a wide range of choices for the materials used to create superinsulated building envelopes for various climates. Wall assemblies can be built using conventional lumber or masonry construction, double-stud construction, structural insulated panels (SIPs), insulated concrete forms (ICFs), truss joist I-beams (TJIs), steel or concrete frame, or straw bale construction.

Similarly, designers can choose from a number of different types of insulation. These include cellulose, high-density blown-in fiberglass, polystyrene, spray foam, and again straw bale. Although spray foams have a high R-value and are easy to apply, many builders prefer not to use them because they are petroleum-based products, have high embodied energy, and because the currently market dominant expansion agents can contribute significantly to global warming. Manufacturers are seeking to develop spray foams that do not have these disadvantages. Vacuum insulated panels (VIPs) are a relatively new, and still pricey, option with an exceptionally high R-value per inch. Using VIPs allows designers and builders to greatly decrease the thickness of the walls in homes where that is a consideration. Still higher-tech insulations are in development such as aero gels.

No matter which type of insulation gets chosen, Passive House builders need to make sure that the product is installed correctly. It is important to assure and measure the proper density on site before insulation is blown in if loose fill insulation such as cellulose or high density fiberglass is being used to prevent any settling. In any case, with or without having measured the installed density during the insulation process, the application and performance of insulation can be directly measured using thermographic imaging. All objects emit infrared (IR) radiation, and the amount of radiation emitted increases with the temperature of the object. Variations in IR radiation, and therefore in temperature, can be observed using a thermographic, or IR, camera – a useful tool for testing buildings during the quality assurance process. Since these cameras can readily detect heat loss, they can usually identify areas where insulation is insufficient, incomplete, damaged, or settled. Technicians who read thermal



images of properly constructed Passive Houses have jokingly called them boring, as they often reveal little substantive heat loss.

### Eliminating Thermal Bridges

Heat loss follows the path of least resistance: Heat will pass very quickly through an element that has a higher thermal conductivity than the surrounding material, forming what is known as a thermal bridge. Thermal bridges can significantly increase heat losses, which can create areas in or on the walls that are cooler than their surroundings. In the worst-case scenario, this can cause moisture problems: when warm, moist air condenses on a cooler surface.

Thermal bridges occur at edges, corners, connections, and penetrations. A bridge can be as simple as a single lintel that has a higher thermal conductivity than the surrounding wall or several steel wall ties that pass through an envelope. A cantilevering balcony slab that is not insulated from, and thus thermally isolated from, an interior concrete floor can be a potent thermal bridge. An effective thermal isolation is called a thermal break. Without a thermal break, the balcony will act as a very large cooling fin – in the wintertime!

In a Passive House, there are few or no thermal bridges. When the thermal bridge coefficient, which is an indicator of the extra heat loss caused by a thermal bridge, is less than 0.01 watts per meter per Kelvin (W/mK) or 0.006 British Thermal Units per hour foot and degree Fahrenheit (BTU/h ft °F), the detail or wall assembly is said to be thermal bridge free. Additional heat loss through this detail is negligible, and interior temperatures are sufficiently stabilized to eliminate moisture problems and to meet the comfort criteria. It is critical for the Passive House designer and builder to plan for reducing or eliminating thermal breaks by limiting penetrations, and by using heat transfer-resistant, thermally broken materials. Here also, thermographic imaging during the quality control visits can be used to determine how effective the efforts to eliminate thermal bridges have been.

### Airtight Construction

Airtight construction helps the performance of a building by reducing or eliminating drafts – whether hot or cold – thereby reducing the need for space

conditioning to maintain comfort. Airtightness also helps to prevent warm, moist air from penetrating the structure, condensing inside the wall, and causing structural damage.

Airtight construction is achieved by wrapping an intact, continuous layer of airtight materials around the entire building envelope. Special care must be taken to ensure continuity of this layer around windows, doors, penetrations, and all joints between the roof, walls, and floors. Insulation materials are generally not airtight and that includes spray foams. The materials used to create an intact airtight layer include various membranes, tapes, plasters, glues, shields, and gaskets. These materials are durable, adherent, easy to apply, and environmentally sound, which in turn makes it easier for a builder to meet the stringent airtightness requirement of the Passive House standard.

### Airtightness of a House: A Measurable Dimension of the Quality of Construction

Testing airtightness requires the use of a blower door, which is essentially a large fan used in conjunction with sensitive measuring instruments. The blower door can be used to either depressurize or pressurize a house to a designated pressure. With the fan set to maintain this designated pressure, a technician can assess how much air is infiltrating the building through all its gaps and cracks. Specific leaks can be detected during the test either by hand, by employing tracer smoke, or by looking at thermographic images if there is a temperature difference between inside and outside. It is best to conduct the blower door test at a point in construction when the airtight layer can still be easily accessed, and any leaks can be readily addressed.

Passive Houses are extremely tight. At a standard test pressure of 50 Pa, a building that meets the Passive House standard must allow no more than 0.6 ACH (*Air Changes per Hour*) in order to achieve certification. Projects that have successfully met the Passive House standard have been built from timber, masonry, prefabricated elements, and steel or concrete frame buildings with superinsulated curtain walls.

The tightness standard of 0.6 ACH50 is not an arbitrary measure. As walls get superinsulated and airtightened, they become more prone to moisture damage. The typical vapor profile and drying potential

changes, sometimes significantly. Consequently, 0.6ACH50 has been determined to be the safe limit for superinsulated walls even in very cold climates to avoid moisture damage through infiltration and diffusion.

Airtightness does not mean that one cannot open the windows. Passive House buildings have fully operable windows, and most are designed to take full advantage of natural ventilation to help maintain comfortable temperatures in the spring, fall, and even the summer, depending on the local climate.

### Heat or Energy Recovery Ventilation

Perhaps the most common misperception regarding Passive Houses concerns air flow. “A house needs to breathe,” builders might say disapprovingly, when first presented with the idea of building very tight homes. Buildings that meet the Passive House standard do breathe – exceptionally well. However, rather than breathing unknown volumes of air through uncontrolled leaks, Passive Houses breathe controlled volumes of air by mechanical ventilation. Mechanical ventilation continuously circulates measured amounts of fresh air through the house and exhausts known quantities of stale air from the house. This makes for excellent IAQ. The amount of air exchanged is strictly based on the exchange necessary to assure that all pollutants get sufficiently removed. Only the air needed for excellent IAQ is exchanged. The health and comfort of the occupants come first for the Passive House designer, and excellent IAQ is indispensable for occupant health.

A Passive House building is ventilated using a balanced mechanical ventilation system. Needless to say, this ventilation system must be extremely energy efficient. To that end, Passive House designers specify energy recovery ventilators (ERVs) or heat recovery ventilators (HRVs) in cold, dry, and marine climates. These machines incorporate an air-to-air energy recovery system, which conserves most of the energy in the exhaust air and transfers it to the incoming fresh air. This significantly reduces the energy needed to heat that incoming air.

State-of-the-art ventilation systems have measured and verified heat recovery rates of 75–92%. The energy consumed by the motor of the ventilator is also

considered. To avoid a total net loss of input energy versus heat recovery effectiveness, the efficiency of the motor needs to be extremely high. The energy efficiency limit for Passive House ventilation system motors should meet 0.45 watt hours per cubic meter ( $\text{Wh/m}^3$ ) of air or 0.76 watts per cubic foot and minute ( $\text{W/cfm}$ ) of air.

The ventilation system generally exhausts air from the rooms that produce moisture and unwanted odors, such as the kitchen and bathrooms. The flow rate is set to a low level. For a typical single family home, the base airflow rate is approximately 90–120 cfm total. Timed overrides are typically installed in the exhaust rooms to allow the user a short period of time to increase the ventilation flow rate if moisture or odor levels are elevated. The exhaust air gets drawn through the ventilator on its way out of the building. There it passes through a heat exchanger that transfers the reusable heat energy to the incoming fresh air. It is important to note that the exhaust air is not mixed with the incoming air; only its heat is transferred. Acceptable contamination of the two air streams is limited to 3% max. While return air is circulated back to the furnace in a forced air system, no air is recirculated with a mechanical ventilation system. All supply air is fresh air.

When operating, the ventilator constantly provides a steady supply of fresh air. At the same time, it removes excess moisture,  $\text{CO}_2$ , other pollutants like VOCs from furniture, unwanted odors and even radon. The incoming air is filtered and balanced. Filtration is important for indoor hygiene as well as for a long-lasting ventilation system. Passive Houses have a requirement of a minimum filter quality of F7 (Europe) or MERV 11. The air is distributed at the low flow rate through small, unobtrusive, but highly effective, diffusers. The system is generally very quiet and draft-free. Dust circulation is minimized. The PHPP recommends an ACH of 0.3–0.4 times the volume of the building, and a guideline ACH of 30 cubic meters ( $\text{m}^3$ ) per person.

The main difference between an HRV and an ERV is that the HRV conserves heat and cooling energy, while the ERV does both and transfers humidity as well. In summer, an ERV helps keep the humidity outside; in winter, it helps prevent indoor air from becoming too dry. For in-between seasons, when no conditioning is needed, a bypass can be installed for either system to

avoid heating the incoming air. Alternatively, the ventilation system can be turned off altogether, and windows can be thrown open to bring in fresh air.

Either system's efficiency can be increased by pre-warming or pre-cooling the incoming air. This is done by passing the incoming air through earth tubes. Since the ground maintains a more consistent temperature throughout the year than the outdoors, passing the air through tubes buried in the earth either pre-heats or pre-cools the air, depending on the season. Pre-heating and pre-cooling can also be accomplished indirectly, by circulating water in an underground pipe and using it to heat or cool the air with a water-to-air heat exchanger integrated with the intake air stream between the ventilator and the envelope. This way, potential condensate in humid climates can drain in a controlled location rather than occurring in the earth tube.

### High-Performance Windows and Doors

In modeling the energy balance of a building, the designers of Passive House buildings choose windows and doors based largely on their insulating value. At one time it was hard to find doors and windows that had the exceptional insulating properties required by the Passive House comfort and the few that existed were very expensive.

That is no longer the case. There have been extraordinary advances in window quality over the past 30 years, and thermal losses from windows have dropped dramatically even for North American products. Many brands of windows and doors are now being made tighter, reducing losses through infiltration and exfiltration. Doors have been provided with appropriate thermal breaks and double gaskets. Overall, high-performance windows and doors are proving to be cost-effective in Passive House applications.

One development that has significantly affected the heat conductivity of window glazing is the introduction of low-emissivity (low-e) coatings. These are microscopically thin, transparent layers of metal or metallic oxide deposited on the surface of the glass. The coated side of the glass faces into the gap between two panes of a glazed assembly. The gap is filled with low-conductivity argon or krypton gas rather than air, greatly reducing the window's radiant heat transfer.

Different low-e coatings have been designed to allow for high, moderate, or low solar gain. This provides a range of options for houses in all climates, from heating dominated to cooling dominated. Today, builders can choose to install triple-pane low-e-coated, argon-filled windows with special low-conductivity spacers and insulated, thermally broken frames. These windows eliminate any perceptible cold radiation or convective cold air flow, even in periods of heavy frost. For the moderate cool climate (climate zone 4) and the central European climate a U-value of 0.85 (W/m<sup>2</sup>K) or 0.15 BTU/h ft<sup>2</sup> °F is recommended for the entire installed assembly accounting for frame values, glass values, spacer and installation thermal bridge effects.

### Optimization of Passive-Solar and Internal Heat Gain

Not only must designers of Passive House buildings minimize energy loss, they must also carefully manage a home's energy gains. The first step in designing a Passive House is to consider how the orientation of a building – and its various parts – will affect its energy losses and gains. There are many issues to be considered. Where should the glazing be to allow for maximum sunlight when sunlight is wanted, and minimal heat gain when heat gain is unwanted? The more direct natural lighting there is, the less energy will be needed to provide light. Designers can enhance residents' enjoyment of available sunlight by orienting bedrooms and living rooms to the south, and putting utility rooms, closets, and circulating spaces, where sunlight is not needed, to the north.

However, it is not always possible to site a house in this ideal way. There may be buildings, trees, or landforms that cast shadows during short winter days, blocking out much of the low sunlight. Or the designer may need to accommodate the homeowner's demand for a certain view – a view that would not be available in an ideal orientation.

Windows are designed, oriented, and installed to take advantage of the outstanding passive solar energy that can be gained through them. But the goal is not simply to allow for as much solar gain as possible. Some early superinsulated buildings suffered from overheating because not enough consideration was given to the amount of solar gain that the house

would experience. A good design should balance solar gain within the home's overall conditioning needs – and within the glazing budget. Even very efficient windows can lose more heat over a year than they gain, depending on their location, and large windows are expensive. In the northern hemisphere, in climates with cold winters, windows on the north allow for no direct solar heat gain, while those on the south allow for a great deal of it. In summertime, and in primarily cooling climates, it is very important to prevent excess solar heat gain. This can be done by shading the windows. Roof eaves of the proper length can effectively shade south-facing windows when the sun is higher in the summer, and still allow for maximum solar heat gain in the winter, when the sun is lower and the days are colder. Deciduous trees or vines on a trellis can also block out sunlight in the summer and admit it in the winter. In climates that have a significant cooling load, the designer should consider limiting unshaded east- and west-facing windows, and specifying only windows that have low-solar-gain, low-e coatings. During the morning and late afternoon, low-angle sunlight can generate a great deal of heat in such windows. A guiding value for Passive House solar gain optimization in colder climates is the recommendation of a solar heat gain coefficient of approximately 50% or slightly higher, in cooling climates it should be below 30%.

Another, perhaps less obvious, source of heat gain is internal. Given the exceptionally low levels of heat loss in a Passive House, heat from internal sources can make quite a difference. Household appliances, electronic equipment, artificial lighting, candles, people – all can have a significant effect on the heat gain in a Passive House. While designers may not be choosing how many or which appliances will be installed in a house, designers often select lighting sources, and must take into account these heat gains when calculating the overall internal energy gains.

### **Energy Balancing – Modeling with PHPP (Passive House Planning Package)**

There are many elements of Passive House design that need to be integrated with one another. They include wall thickness, R- or U-values, thermal bridges, airtightness, ventilation sizing, windows, solar orientation, climate, and energy gains, and losses. Modeling

software for Passive Houses needs to facilitate accurate energy use prediction helping a designer to integrate each of these components into the design so that the final design will meet Passive House requirements and the projected performance. One starts with considering the whole building as one zone for the energy calculation. The designer considers all of the house's basic characteristics, including orientation, size, window location, insulation levels, and so on. One then computes the energy balance of the design and calculates the equivalent “miles per gallon” energy consumption for the house. If needed, the designer can change a house's components – window location or size for example – and model the impact of those changes on the overall energy balance.

Out of all energy balancing models out there, the PHPP, Passive House Planning Package, developed by the Passivhaus Institut in Germany, is a very responsive what-if tool, allowing designers to readily shift the variables of design to reach these goals. It also effectively models such things as solar water heating for combined space heating and domestic hot water, natural ventilation such as night cooling, and the efficiency of heat/energy recovery ventilation. The PHPP incorporates an impressive depth and level of detail when considering the variables that create a building's unique energy balance. Other modeling tools on the market are starting to incorporate similar strategies increasing their calculation granularity to calculate the energy balance of extremely low energy buildings accurately and could be used to calculate Passive Houses as well.

**A Word on Cooling and Dehumidification** The Passive House Building Energy Standard was developed primarily in Central Europe, which has a relatively mild, primarily heating-oriented climate. Implementation of designs that meet the Passive House standard is more challenging in climates of more extreme cold, heat, or humidity. Already, many Passive Houses have been built in extremely cold climates, a few Passive Houses have been completed in hotter climates and more are being submitted for certification review.

In cooling load (as in heating load) situations, the space conditioning load must be minimized. This takes careful planning. As explained earlier, the high levels of insulation in a Passive House help to keep indoor

temperatures cool. In addition to the standard measures for preventing excess solar gain, convective venting behind siding and roofing and night cooling will often help to maintain indoor comfort. In humid climates, an additional cooling load may stem from the need to remove latent moisture from the air. A very small and efficient air-to-air heat pump – also known as a mini-split – can remove this moisture and provide adequate cooling. In extremely humid climates, such as in Louisiana, additional dehumidification might be necessary.

**Economic Sustainability** Passive House design focuses on balancing energy gains and losses in order to attain a level of energy efficiency that is far beyond the norm. But the norm is changing, and many people now recognize that energy efficiency is profoundly important, both economically and environmentally. It has been called a low-hanging fruit, an innovation that can, and should, be readily attained.

This focus on energy efficiency makes Passive Houses more expensive to build. Construction costs generally run 10–15% higher for single family homes than the costs for conventional houses. The additional upfront costs for more insulation, better windows and doors, and more labor for higher quality installations are partially offset by the lower cost of the heating and cooling systems. Because Passive Houses have such small heating and cooling requirements, conventional heating and cooling systems can be replaced with miniaturized components and efficient mechanical ventilation. This is one example of how the integrated planning required to build a Passive House helps builders to “tunnel through the cost barrier,” in the words of Amory Lovins. The additional construction cost is readily recovered in savings on the homeowner’s energy bill. These savings will continue throughout the life of the house. And – perhaps most important in the long run – the Passive House will generate a carbon footprint that is a fraction of the size of that of a conventional house. On-site renewable energy resources can be added to create a true zero-energy, or even plus-energy, house – one that produces more power than it consumes.

In the past, most homes were built with scant attention paid to their long-term energy consumption. This approach needs to change. It is important to use our

limited natural resources wisely and to build our homes with quality and durability in mind. The costs of energy consumption are high, and they are going higher. The savings to be realized over the life of a Passive House are remarkable, both economically and environmentally.

### **Future Directions of Passive House in the USA, Canada, and Internationally**

The Passive House Institute US was the first outpost in North America that made it its mission as a nonprofit organization in 2007 to increase the number of ultra energy efficient buildings in the USA by promoting the adoption rate of the Passive House Building Energy Standard. The organization has worked diligently with limited capacities to provide services such as Certification, Design, Training, Education, and Research nationwide. The National Passive House Alliance, a nationwide membership organization focuses on educating its members and the general public on the topic, improving the recognition of Passive House construction and principles in various legislative programs, and providing continuing education opportunities to its members. Since 2008, the number of Certified Passive House Consultants in the USA and Canada has been exponentially growing. The demand is expected to increase significantly over the next years. Passive House buildings in the certification process in the USA are now representing commercial and residential buildings in almost all NA climate zones. This number also continues to grow rapidly. Now, also other organizations and the government have become interested, such as the USGBC as well as federal and state legislators, to include Passive House as a Low/Net Zero Energy alternative into their programs.

The success and fast uptake over the past 4 years of a performance-based energy metric clearly defining the level of conservation that should be reached was surprising. It has made it clear that there is the need in the market to increase capacity to train professionals as well as students in the field of Passive House design and construction. This not only helps Passive House to meet environmental needs but it is also a way to reclaim many jobs in the construction industry that were lost in the housing melt down.

While the initial success in North America is hopeful it is also a reason to be cautious. There is an

increased need for the output of quality publications and design guidelines documenting especially the experiences of Passive House designs in North American climates to date and to increase monitoring programs of larger developments in various climates that will serve as proof of concept. In conclusion, it can be said that the Passivhaus standard, as it is known today in central Europe, was developed in a relatively easy and homogenous climate, a climate most closely resembling the climate in the North West of the USA. The recent Passive House iterations of the USA and Canadian projects have shown that still a lot of research and transfer is needed to assure an equally successful and widespread dissemination of the concept, especially in the very cold and humid, mixed humid, and hot humid climates of the USA where the energy savings potentials from conservation are extremely high. It is to be expected that there will be variations on the theme from the central Europe so successful core Passive House metric developed in Germany. Here, in North America, the movement is just beginning its way to Main Street.

Internationally, the implementation of the principles in different countries poses new challenges altogether. Existing building culture, exiting building components in the market, and cultural acceptance of the Passive Design standards will play a big role in how far the concept can continue to grow internationally also. Many initial groups have formed from New Zealand to China and Russia disseminating the knowledge similarly to what PHIUS has done in the USA over the past 6 years. The International Energy Agency, headquartered in Paris, France, has invested a lot into research on how to guide and make recommendations to governments with regard to the implementation of highly efficient building codes aiming at envisioning cost-effective zero energy buildings by 2030 as an international goal. The successful strategy will most likely be a coordinated guided code implementation approach coupled with a grass roots effort from organizations and individuals working from the bottom up.

### For More Information

The Passive House Institute United States (PHIUS) is a nonprofit certifying, consulting, and research firm working to further the implementation

of Passive House Building Energy Standards nationwide. For information about the institute, go to [www.passivehouse.us](http://www.passivehouse.us).

## Passive Solar Heating in Built Environment

ROBERT HASTINGS

AEU Ltd. Architecture, Energy & Environment,  
Wallisellen, Switzerland

### Article Outline

Glossary  
 Definition and Importance of Passive Solar Heating  
 Introduction  
 History  
 Principles, Applications, and Integration  
 Direct Gain  
 Indirect Gain  
 Isolated Gain/Hybrid  
 Future Directions  
 Acknowledgments  
 Bibliography

### Glossary

**Solar architecture** The deliberate use of solar energy by means of the building architecture, thereby reducing purchased energy dependence while enhancing the quality of enclosed space.

**Passive** Not requiring actions to achieve a desired goal. In the case of passive solar energy use, solar energy is captured and distributed in a building without machinery by using the physics of conduction, free convection, and radiation.

**Direct gain** The direct gain of heat within a building by sunlight entering through glazed openings in the enclosure, which then traps and stores the heat.

**Indirect gain** Solar energy absorbed in some fashion on or in walls or roofs and converted to heat. This heat either remains entrapped in the building envelope to reduce building heat losses, or it is transferred into the building by conduction or convection. There may be a delay between the

time when sunlight is absorbed and when heat penetrates into the enclosed volume.

**Isolated gain** Solar energy absorbed outside the insulated building envelope and then transported by free convection to the enclosed volume.

**Solar air system** Type of isolated gain system where heat from the collector transported to the point of use or storage by air (verses water in active thermal systems).

**Hybrid solar system** A passive system assisted by a small fan to increase system efficiency, possibly PV-powered. The energy ratio of heat output to electrical input can easily exceed 20:1.

### Definition and Importance of Passive Solar Heating

Passive solar heating is the use of solar energy to heat a building without mechanical or electrical energy. The architecture and construction capture, store, and distribute the sun's energy. Every building with windows exposed to the sun is passively heated, but heat losses may exceed the solar gains. Accordingly, if the passive heat gain is to reduce heating costs, the system heat losses must be minimized. Ideally, the concept includes mass to store daytime solar heat for nights, increasing the usability of the gains. Finally, the heating system must shut off when solar heating achieves the desired room temperature. Two constraints on passive solar use are glare control and shading during non-heating months.

Maximizing usable passive solar gains is an important design aspect, but often designers focus only on minimizing heat losses. Taking the finance world as an analogy, no one will accumulate wealth through savings alone, income must be maximized and wisely invested. So, not only reducing heat loss is essential to low energy architecture, maximizing solar gains is important, as it has been over the history of building.

An often cited early example of solar design awareness is the "Megaron House" described by Socrates in the year 400 B.C. Numerous other examples can be found, i.e., the New England "salt box" of the seventeenth century or Swiss farm houses of the eighteenth century. In the twentieth century, the term "solar house" became popular and following the first oil

shock of 1973, the term "passive solar buildings" was coined. In all these examples, the basic principles are the same; maximize the south exposure of a building to capture as much solar heat as possible and insulate the enclosure well to keep the heat in.

Passive solar building design is an important means for slowing climate change by reducing the burning of fossil fuels. It is not, however, a least first-cost way to build. Larger, better insulating windows or opaque collector constructions cost more than conventional constructions. Three arguments justifying this investment are:

- Long-term (>10 years) good return on the investment
- Economy and security as future fossil fuel prices increase and supply subject to interruptions
- Living qualities of passive solar buildings flooded with light and natural warmth (Fig. 1)



**Passive Solar Heating in Built Environment. Figure 1** A living room flooded with sunlight from large south-facing windows (photo source: robert.hastings@aeu.ch)

## Introduction

### Concepts

Buildings that consume less fossil fuel are “nice to have” today, but will be essential in the future. Since buildings are long-term investments, they must be built or rebuilt looking to the future.

Heating is a major use of energy in buildings and dependence on fossil fuels for heating can be dramatically reduced through two strategies: reducing heat losses and increasing the use of solar heat. Logically, a combination of these two strategies is desirable.

Solar energy can be used by passive or active means:

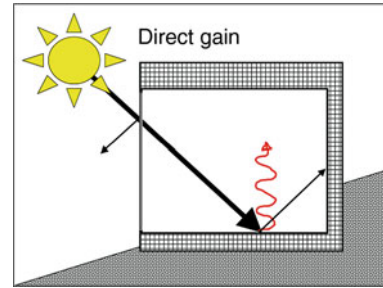
- *Passive solar use* does not rely on mechanical components to capture, store, and distribute the heat, the building construction fulfills these functions.
- *Active solar use* typically involves a remote solar collector and a pump or fan to transfer the heat to storage and from storage to point of use.

A low energy building must lose as little heat as possible, hence the importance given to insulation, air tightness, and heat recovery. An example design standard promoting extreme energy conservation is the “Passive House Standard” [1]. To meet this standard in Europe, three requirements must be fulfilled:

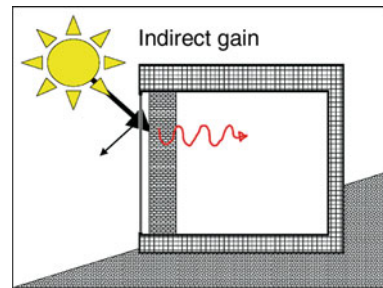
- The annual heating requirement must be less than 15 kWh/(m<sup>2</sup>a) or maximum heating power 10 W/m<sup>2</sup>a based on the net heated floor area.
- The combined primary energy consumption for heating, hot water and household electricity may not exceed 120 kWh/(m<sup>2</sup>a).
- The air leakage of the enclosure tested under a pressured difference of 50 Pa ( $n_{50}$ ) may not exceed 0.6 house air volumes per hour.

A passive solar building is not defined to this extent; it simply describes a structure in which the designer deliberately maximized using solar energy passively. So, in fact, a passive house can also be a passive solar house and indeed, in the planning recommendation for a passive house, using passive solar energy is encouraged and credited.

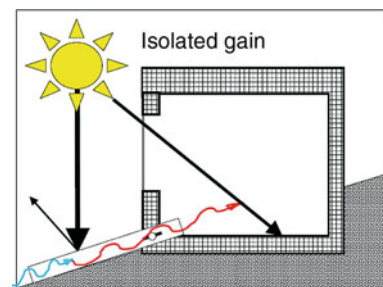
Three passive solar heating concepts were defined after the first oil shock of 1973 and are still useful today:



*Direct gain:* Windows capture the sun in a well-insulated building; interior construction mass stores the potentially excess daytime gains into the night; and some form of shading provides comfort during non-heating seasons. Direct gain is the oldest and still most cost-effective concept, given its potential to also enhancing the quality of life in buildings.



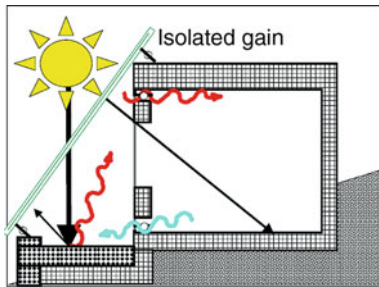
*Indirect gain:* The building envelope captures solar heat, which is then conducted and/or convected to the building interior, possibly with a time delay of up to 8 h. Alternatively, the goal may be simply to capture enough solar heat in the envelope construction to eliminate heat losses from the building much of the time, i.e., a dynamic  $U$ -value over the heating season approaching zero. Indirect gain systems nicely compliment direct gain systems.



*Isolated gain passive:* Solar energy is converted to heat outside the insulated building envelope and then



delivered to the building interior or storage. This can be by gravity-driven convection, or with the help of a small fan (a “hybrid” system). While not purely passive, hybrid systems are reported here because the proportion of delivered heat to electrical energy is so small.



A sunspace or attached greenhouse with controlled opening to the building is also an isolated gain system. Isolated gain systems are the most complex and expensive, but offer the most control of when and how much solar heat is delivered into the building.

It can be useful to consider passive solar heating opportunities by building types and climates. Note, that in this section, locations north of the equator are assumed. South of the equator, north orientations have priority.

### Building Types

Buildings where heating loads dominate over cooling loads are the obvious candidates for passive solar design, i.e., residential buildings and small commercial or institutional buildings. Three factors are decisive here:

- As the ratio of enclosing surface to enclosed volume increases, heat loss increases, so a solar heating can be more beneficial.
- As the density of heat production from people or appliances increases, the usefulness of solar heat decreases.
- Direct solar gains in the form of heat and light are a combined asset, i.e., for hospitals, old age homes, and schools as well as residences. (Examples and Design insights for passive solar use in commercial and institutional buildings were researched and documented in an IEA project [2].

The energy optimization of a building must balance passive solar and daylight benefits against mechanical cooling and electric lighting energy demands, both of which have very high primary energy factors.

### Climates

*Northern climates* such as Scandinavia would seem to pose a problem for passive solar use. Winter days are short, the sun is weak, and the sun path is at a very low angle. This means, however, that windows or vertical collection surfaces intercept the sun at a more direct angle. Furthermore, the heating season is very long, extending from early autumn to late spring. Before 21 September or after 21 March, heating is still needed, when days are longer than in southern latitudes. Passive solar concepts must maximize the usefulness of spring and autumn solar heating, while minimizing mid-winter heat losses.

*Temperate climates* are the ideal situation for passive solar buildings. Not just sunny temperate, but also overcast temperate climates. This has become possible with the development of very high-performance glass ( $U$ -value  $< 1.0 \text{ W/m}^2\text{K}$ ). Consider the example of diffuse solar radiation at only  $100 \text{ W/m}^2$  for 6 h and an ambient temperature of  $5^\circ\text{C}$ . The solar gains through a glass with a  $g$ -value of 0.5 (admitting 50% of the solar radiation) will offset the 24 h heat losses of a glass with a  $U_{\text{glass}}$  of  $0.8 \text{ W/m}^2\text{K}$ . If the sun shines with more intensity or more hours, it is a passive solar winner. Because temperate climates often have hot summers, the concept must also include shading.

*Mild climates* offer a challenge: to achieve zero-heating energy buildings by combining passive solar design and conservation without degrading summer comfort. This is at least as challenging as achieving net-zero-energy buildings. The latter achieve a net zero balance by taking a credit from the summer electrical output of a large PV-roof (multiplied by a high primary energy factor) against the energy deficit in winter, which must somehow be met. Passive solar heating of a highly insulated building can answer part of the “somehow” question.

## Strengths and Weaknesses

+	Living quality: daylight and naturally warmth from the sun's warmth.
+	Security: in the event of energy supply interruptions.
+	Costs: only the marginal costs of added aperture area, be it collector or window area and mass, must be amortized by energy savings.
+	Return on investment: As energy prices rise, return on the investment in passive solar measures increasingly attractive.
+	Low maintenance: There are no maintenance costs for pumps or fans.
+	All of these factors can positively affect resale value of the property.
-	Operation: often passive solar use requires active occupants adjusting sun-shading or opening windows or vents.
-	Poorly designed or incorrectly used passive solar buildings may use more energy than conventional buildings. Informed design, strict quality control, and intelligent operation are essential.

## Road Map to This Section

A historic review of passive solar design shows how this approach has developed in parallel with technological developments of building components. It is instructive to examine which concepts came into existence, evolved, flourished, or died out. This may save reinventing a broken wheel, or ideas might arise for new variations or concepts.

*Principles and applications* review different approaches to passively capturing, storing, and using solar energy to heat buildings.

*Direct, indirect, and isolated gain* concepts are reported in detail.

*Finally, the past, present, and future* of passive solar heating are discussed in the context of expected energy supply developments, demographics, and increasingly well-insulated and automated buildings.

## History

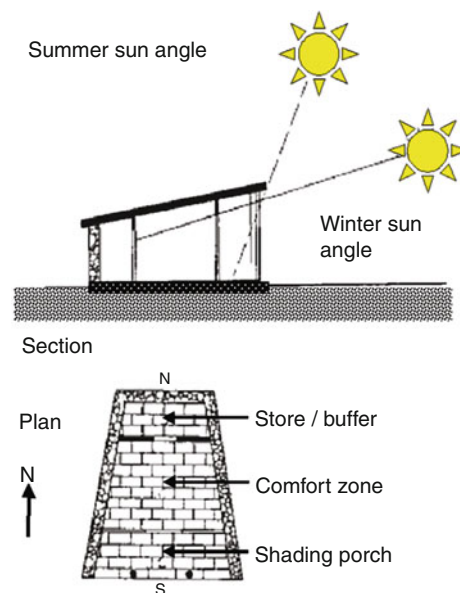
Concepts for passive solar heating date back millennia. Materials and components were very primitive by today's standards, but comfort expectations were also

much lower. A net solar gain is possible even with single glazing if the required room temperature is only 16°C. The twentieth century saw dramatic developments in material science and production techniques, e.g., in glass production. The evolution in the last century has been equally spectacular. Single glazing at the beginning of the twentieth century ( $U = 5.8 \text{ W/m}^2\text{K}$ ), evolved to fused double glazing in the 1950s ( $U = 2.8 \text{ W/m}^2\text{K}$ ). Insulating glazing ( $U = 1.2 \text{ W/m}^2\text{K}$ ) in the 1990s is now available in triple glazing ( $U = 0.5 \text{ W/m}^2\text{K}$ ), or more than a factor 10 better than window glazing a century ago.

As a result, some concepts, which earlier proved ineffective for a given climate or building type, may indeed be effective today and should be "rediscovered." Following is a short-time journey through the evolution of passive solar heating.

## Ancient times

The most often cited example of awareness of passive solar use is a concept house, the "Megaron House" (Fig. 2) described by Socrates (469–397 B.C.). He expressed the following thoughts: "Doesn't the sun shine into houses facing south in winter, whereas in summer the sun wanders over us and the roof so that we have shade? Because this is comfortable, then

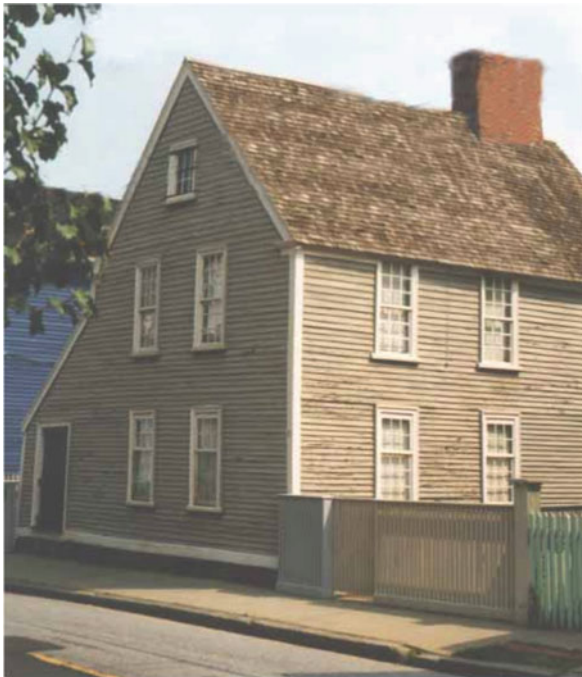


**Passive Solar Heating in Built Environment. Figure 2**  
The Megaron House concept described by Socrates

south-oriented rooms should be built higher in order not to shut out the sun, whereas the north rooms should be lower because of the cold north wind.” This was the logic for this funnel-shaped house concept, opening in plan and section to the south. A roofed porch admitted sunlight into the main room in winter but shaded it in summer. A room to the north served as both storage and as a buffer from the north exposure.

### 1600–1900

*New England Salt Box:* A classic passive solar house form appeared between 1650 and 1830 in New England, the “Salt Box” (Fig. 3). Its name is derived from the shape of boxes used to store salt at that time. Initially, the house form came about when an addition was made to the rear and the roof slope carried down from the two-story main house. Typically, the addition incorporated a kitchen with its own fireplace, a pantry, and a room for child birth or nursing the ill [3]. The main chimney rose inside the house to keep its heat inside. Also, very practical are the double-hung windows.



Passive Solar Heating in Built Environment. Figure 3 A New England “salt box” house with large south façade and long protecting roof to the north (photo source: robert.hastings@aeu.ch)

The sashes were hung on ropes with counterweights of iron or bricks in a cavity of the window frame. The upper sash could be lowered and/or the lower sash raised independently. The height difference between the upper and lower openings induced air circulation.

*Swiss Appenzell House:* The Swiss Appenzell houses from the eighteenth to early nineteenth century had facades with many window bands protected by a projecting roof at each storey (Fig. 4). This afforded summer shading and weather protection for the windows. The curved white under surfaces captured and deflected additional daylight down to the windows.

### 1900–1950s

In The year 1927 saw a breakthrough in glass production. Using the Penn verson Drawing Machine, glass was pulled through rollers in a new process implemented by PPG Industries. For the first time,



Passive Solar Heating in Built Environment. Figure 4 An Appenzell house with large window area protected by multiple roof projections (photo: robert.hastings@aeu.ch)

large sheets of glass could be produced. This opened exciting new architectural possibilities, but with large heat losses and comfort problems. With the introduction of insulating glass by LOF, it was possible to have large window areas and net solar heat gains. Architects played with the design opportunities this new technology offered. Researchers quantified how long a room could be kept warm by what outside conditions. The press publicized what was then possible with new solar houses. Solar buildings were a mainstream topic. An example of such architecture is the work of the Architect George Fred Keck. [Figure 5](#) shows the living room with a stone floor and fireplace to absorb the sunlight flooding in from the full southwest front of windows [4]. This house, built for Dr. and Mrs. Hugh Duncan of Flossmore, IL, United States, was monitored by two researchers. The performance was surprisingly good. One winter day in 1941, when the ambient temperature was  $-20^{\circ}\text{C}$ , the heating system shut off by 08:30 h and stayed off until 20:30 h [5].

### 1960s

Oil was plentiful and cheap, everyone was happy, renewable energy was not a topic of any popular importance and very little happened.

### 1970–1980s

In 1973, an oil embargo imposed on the United States led to a crisis of historic proportions. Americans can react astonishingly effectively and quickly to a crisis



**Passive Solar Heating in Built Environment. Figure 5** Direct gain maximized in the Duncan House (picture by permission of Pilkington, North America, Inc.)

and this was the case then: “overnight,” a national program to reduce foreign oil dependency was initiated. The Energy Research and Development Administration (ERDA) was activated on 19 January 1975 and the Solar Energy Research Institute (SERI) in Golden CO was founded. The department of Housing and Urban Development (HUD) held a national competition with grants for building the solar houses. Many built projects were monitored by national laboratories and published [6], as, for example, the Balcomb house shown in [Fig. 6](#). In 1977, the first National Passive Solar conference was held [7] and in subsequent years, each conference drew over a 1,000 enthusiasts.

Passive solar use was a major topic of the American Solar Energy Society (ASES), a national organization linked with the International Solar Energy Society (ISES). These were the boom years for passive solar buildings. Research and demonstration projects were well funded at the federal and state levels. Atomic physicists “saw the light” and became solar building physicists at renowned national laboratories, including Los Alamos, Lawrence Berkeley, Brookhaven, and the National Bureau of Standards. Exemplary demonstration projects were sponsored by the Tennessee Valley Authority (TVA), an enormous interstate electrical utility. Regional solar energy centers oversaw the evaluation and publicizing of countless solar buildings.

To help energy consultants, researchers, and academics analyze concepts, complex computer models were developed. These could quantify the dynamics of solar and heating input, heat storage, and building heat



**Passive Solar Heating in Built Environment. Figure 6** The Balcomb house in New Mexico (photo source: robert.hastings@aeu.ch)

losses. Auxiliary heat demand and comfort performance were reported on an hourly basis. These tools were, however, difficult to use. Input was cumbersome and errors easily made. Computers in the 1970s still had to be “spoken to” via punched cards. The input was in rows of numbers, separated by spaces or commas punched into cards. Examples of programs include DEROB, NBSLD and BLAST, and later, DOE2. To provide design consultants (designers couldn’t compute) with calculation tools, two approaches were followed:

- Gigantic data bases were computed using research computer models for all thinkable design variations, and then clever nomographs generated. The *Passive Solar Handbook* by Doug Balcomb and R. Jones is a classic example [8].
- Simplified calculation tools were programmed, such as SERIRES (later called SUNREL) and CALPAS. These were small enough to run on the first versions of portable computers (“mini” or “midi” computers).

The goal was to learn how sensitive performance was to a given parameter. To demonstrate how terrific a design was, it was useful to compare it to a conventional builder house of the time. For this purpose three reference houses were defined, based on statistics from the national home builders association (NHBA). The reference designs were published by the National Bureau of Standards (NBS, today NIST) [9].

To be sure, the computer models were telling the truth, measurement data from components and even whole buildings were essential. Test cabins for monitoring systems became a common sight at many national research facilities. [Figure 7](#) shows a test house with an interchangeable modular south façade and clerestory windows sun lighting the north rooms.

Meanwhile at universities, architecture schools continued to teach Le Corbusier as the model for good design. Energy and solar use were not significant design issues, with the exceptions of a small but growing number of architecture and engineering professors. They were the authors of some superb text books, which clearly presented passive solar design principles. Some examples are a passive solar textbook for architects [10], a guide for adapting solar concepts to regional climates and constructions across the whole



**Passive Solar Heating in Built Environment. Figure 7** NBS (NIST) test house 1980, Gaithersburg, MD (architect and photo source: robert.hastings@aeu.ch)



**Passive Solar Heating in Built Environment. Figure 8** The Michel-Trombe wall house in Odello, FR (photo: robert.hastings@aeu.ch)

continent [11], and guidelines for window design strategies to conserve energy [12].

During this period, there were a few good examples of passive solar innovation in Europe as well. The Michelle-Trombe Wall concept was a notable example ([Fig. 8](#)). The original pilot building was constructed at

the Centre National de la Recherche Scientifique (CNRS) in 1967 in the south of France and further developed with a vented version of the wall in 1974 [13].

### 1990s

Europeans began to take interest in the American passive solar movement. Many architects and building researchers travelled to the United States to personally visit passive solar houses. Passive houses began to appear across Europe, from Scandinavia to Italy. National research programs investigated how to optimize passive solar concepts to local European climates and constructions. This was essential. Several passive solar buildings did not function as hoped. Europe gets less sun than New Mexico!

During this time, windows were still mostly double glazed or at best triple glazed ( $U_{\text{glass}}=3.0$  or  $2.2 \text{ W/m}^2\text{K}$ ). Glazing with selective coatings and noble gas fillings were just beginning to enter the market. Accordingly, only windows facing south achieved a net passive solar gain. In northern climates, night insulation of windows was needed for the long dark winters. Several innovative, but expensive roll-down insulating blankets were developed for windows. These largely disappeared from the market as high-performance glazings appeared.

By the end of the 1990s, the growing pains of adapting passive solar architecture to European climates and constructions were over and countless exemplary projects had been built and published. An IEA SHC program searched out and documented exemplary projects [14].

### 2000

During this period, many conventional passive solar-heated houses were built across Europe. Sunspaces were a favored architectural element. Many houses included active solar systems to heat domestic hot water, with Austria leading in the number of such houses. European architects often succeeded in adapting passive solar house designs into good architecture. An example project from 1992 by a Norwegian architect practicing in Austria, Sture Larsen is shown in Fig. 9 [15]. The exterior of the house is in light, wooden framing, the interior is in massive construction. Solar heated air is



**Passive Solar Heating in Built Environment. Figure 9** An Austrian passive solar house with solar air radiant heating and a sunspace in Nüziders, Vorarlberg (architect and photo source: Sture Larsen, [www.solarsen.com](http://www.solarsen.com))

circulated through the walls and floors to radiate into the rooms. A sunspace also helps heat the house.

By the year 2000, a new concept, The Passivhaus (Passive House) had become well established and on the way to becoming the new mark of excellence in low-energy design. It came out of the PHD work of Wolfgang Feist under his Professor, Bo Adamson in Sweden. To reach this standard requires a highly insulated, thermal bridge-free, and airtight building enclosure. Mechanical ventilation is needed to assure good air quality by such tight construction. Heat from exhaust air is then recovered to preheat incoming air. The ventilation air can be used to deliver the small amount of heating needed (maximum  $15 \text{ kWh/m}^2$  heated floor area). Obviously, passive solar heating of such houses is also desired, but challenging to dimension because of the small heating load and short heating season.

### 2010

Today, in the second decade of the new millennium, the term “passive solar heating” is less common. This is paradoxical because with new window frame and glazing systems, highly insulated building envelopes, and sophisticated heating control systems, passive solar gains can cover all heating needs for an extended part of the year in temperate climates. However, the interaction between passive solar gains, internal gains and envelope heat loss needs to be carefully studied to assure comfort and the hope for energy savings.

The former research computer models to study passive solar building concepts, requiring several hours on a main frame computer, today can run in seconds on a laptop. However, today's tools do not consider many of the phenomena the former models did, such as when mass is directly sunlit or only indirectly warmed by room air, or how passive solar heat in south rooms convects to other rooms. This can strongly affect passive solar usability and comfort.

## Principles, Applications, and Integration

### Principles

Passive solar heating requires glass, frames, seasonal sun shading, mass, and extra planning effort. The economics are clear: energy won is more expensive than the energy saved by adding insulation or eliminating air leakage up to a certain point. However, as the insulation thickness is increased, the marginal energy and economic benefits of the next increment decrease. Further, when conventional thicknesses are exceeded, the costs of anchoring the insulation and detailing jump. In contrast to this, the cost of a larger window is not proportional to the window area increase. Larger windows lose less heat per unit area. There is less perimeter for the glass area, so edge losses are smaller. Also, there are increased benefits such as more daylight, the view outside and sense of well being from being sun-warmed (especially for cats). The challenging questions are therefore, which passive solar heating concept is most effective for a given building type and climate, and how big should the system be?

### Applications by Building Types

**Well-Suited Building Types** *Residences* are the most common passive solar application. Detached single-family houses with four outside walls, a roof, and earth contact can best benefit from passive solar gains. Row houses and apartment buildings have many units with only two exposures. If they face east and west, passive solar heating is difficult. One idea would be to install an indirect gain system on the south-facing end walls to compliment morning and afternoon direct gains.

*Large buildings* suitable for passive solar indirect heating include warehouses, gyms (American), or

athletic halls (i.e., tennis halls) [2]. Because a lower air temperature is acceptable or even desired, passive solar gains can make a greater contribution to meeting the heating demand. Heat losses of glazed areas decrease proportionately with less inside to outside temperature difference. Lower required space temperatures increase system efficiency and number of hours when useful passive solar heat can be delivered.

*Swimming pool halls* are a potentially good building type for passive solar heating because a high air temperature is needed, so there is a very long heating season, well into long day spring and fall seasons. Also, there is a great appeal for the space being sunlit. Direct gain, indirect gain for radiant comfort, and isolated solar air systems for humidity control are possibilities.

**Limited Cases** *School* class rooms have high internal gains and high ventilation requirements. The benefit of passive solar heating occurs primarily during heating season weekends and holidays. At that time, typically there is a temperature set back and temperature swings are tolerated, maximizing the usability of passive solar gains. The obvious choice is direct gain with daylighting within the constraints of glare control, thermal comfort near the windows, and the view out being more interesting than the view to the front. Isolated passive solar heating is another alternative. Figure 10 shows a Swiss school with glazed-in balconies off the classrooms.



**Passive Solar Heating in Built Environment. Figure 10** A Swiss school in Gumpenwiese ZH with sunspaces tied into the ventilation concept [16] (photo: robert.hastings@aeu.ch)

Sunspace heated air preheats incoming ventilation air over a heat exchanger for the classrooms [16].

*Old-age homes, nursing homes, and hospitals* can benefit especially from direct gain with daylighting. Indirect gain systems to supply heat at night or isolated gain systems to preheat ventilation supply air are further possibilities. The occupancy tends to be “24/7” and demand somewhat higher room temperatures, extending the heating season.

*Hotels* have a transient occupancy. Guest rooms may be vacant with no internal loads for heating, but must be kept at room temperature or, if the thermostat is set back, very quickly warmed up. Accordingly, passive solar gains to maintain a minimal room temperature can be very useful, but mass can slow the heat-up. Daylight and view out may be assets, but overheating is totally unacceptable. Hotels in cities often may have to be isolated from traffic or airport noise. This can be solved with acoustical glazing or by using an indirect or isolated passive solar concept.

**New Construction Verses Renovation** *New construction* should have very low energy demand as a result of very good insulation, mechanical ventilation with heat recovery. They likely will have a sophisticated heat production, delivery, and control system. Therefore, internal heat from occupancy and appliances will maintain the desired room temperature later in autumn and earlier in spring. The design of passive solar heating must consider this shorter heating season. Storing solar heat is very important because the gains can quickly exceed demand and result in overheating. Ideally, the thermal mass should be sunlit directly. Indirect passive solar systems can contribute to helping shorten the heating season.

*Renovation* is an excellent opportunity to increase passive solar heating. Older buildings have a greater and longer heating demand than well-done new structures. The subject of renovating existing housing was studied in a 4 year project of the Solar Heating and Cooling Program (SHC) of the International Energy Agency (IEA). As part of this work, 60 exemplary projects and an overview with insights were documented in brochures and are available on the Internet (Fig. 11) [17]. Included are apartment buildings, row houses, and single-family houses as well as the special case of



**Passive Solar Heating in Built Environment.** Figure 11 Renovation with solar and conservation, 60 examples of projects across Europe and Canada [17]

historic buildings. The examples come from 10 countries: AT, BE, CA, CH, DE, DK, I, NL, NO, and SE.

**Inappropriate Building Types** Since the sun shines during the daytime, all buildings that do not need heat during the day are not good candidates. Large office buildings, shopping centers, airport terminals are examples of inappropriate building types. Such buildings must cope with energy-intensive cooling problems.

### Applications by Climate

The best climates for passive solar heating buildings are climates that are sunny (so there is energy available) and cold (so there is heat demand). Buildings located at high elevations often have both sun and long heating seasons, which is ideal for passive solar heating. Not surprisingly, there are many good examples of buildings in alpine regions in Europe or the Rocky Mountains in the United States. The absolute best climate for passive solar heating is the southwest of the United States, which is why the passive solar renaissance after the oil shock began there. A very good tool for generating climate data to analyze systems is Meteonorm [18].



*Cold climates* in northern regions have short days in mid-winter. This disadvantage is somewhat offset by the heating season beginning before the autumn equinox and extending past the spring equinox. At those times, days are longer than in more southern latitudes. A further help is that, due to the low sun path, windows or wall collectors intercept sunrays more directly.

*Temperate climates* are well suited for all types of passive solar heating, as is evidenced by the many examples in temperate regions of North America and Europe. In overcast, temperate climates, direct gains systems can profit from even diffuse solar gains, given the highly insulating glass available today.

*Mild climates* pose the challenge to achieve zero heating energy performance through conservation and passive solar measures. Paradoxically, people in mild and sunny climates have the least interest in passive solar design. This is perhaps due to priority being given to passive cooling.

*Dry climates* generally have very good solar availability and large day–night temperature swings. Passive solar heating with mass for storage can be very effective.

*Humid climates* are a problem for both passive solar heating and natural cooling. Humidity reduces solar intensity, day–night ambient temperature swings and blocks night sky radiation for natural cooling. Humid climates are difficult.

### Architectural Integration

Direct gain, indirect gain, and isolated gain are simple concepts; the challenge is to translate a diagram into architecture. The aesthetics of solar design is interesting to observe historically.

Before the twentieth century, building glass was expensive. In some cities, windows were even taxed. Because of their value, they were carefully and artistically integrated into a façade design. Baroque facades are a beautiful example of the celebration of windows, in contrast to the austere holes punched in “contemporary” buildings, as can be seen in a street photo taken in Pilsen CR (Fig. 12).

Early twentieth century architects could play with large glass formats for the first time, but the resulting architecture, while making history books, often resulted in buildings that were an energy disaster.



**Passive Solar Heating in Built Environment. Figure 12** Classical and “modern” fenestration of facades, two contrasting buildings in Pilsen CR (photo source: robert.hastings@aeu.ch)

After the oil shock of 1973 passive solar design (post oil shock), an epoch of innovation and experimentation, both technically and also aesthetically, began. The aesthetics varied greatly, from “California hippy” and New Mexico oil drum to developer Colonial Style. Inventors developed movable reflectors to concentrate sunlight on passive solar elements and movable shading systems for overheat protection. Also, thermal mass became a design opportunity, with glass blocks filled with colored water, glazing filled with phase change material, and rock bins as standing wall elements. Many components were used before they were technically mature. As a result, some components disappeared as quickly as they had appeared, and a design aesthetic never really matured.

By the end of the twentieth century, only technically and economically viable passive solar heating concepts remained. In Europe, the aesthetics of passive solar architecture profited from the attention given to detailing and superb craftsmanship. Design also suffered, however, from the box form architecture which became a craze and results in sterile, boring cubes. Name architects began to apply such concepts, wanting to profit from the growing environmental awareness. North American architects shifted to green architecture, with an emphasis on use of natural material and renewable

energy in environmentally benign designs. The best examples are corporate offices and local institutional buildings from schools to park headquarters.

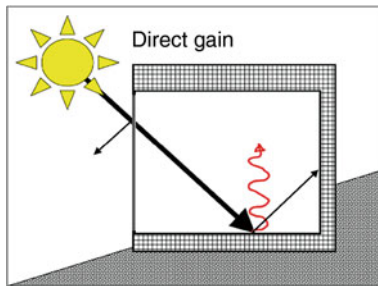
In the twenty-first century, the focus is on conservation. Manufacturers have responded to demand from Passive House planners, so there are now very good components on the market. An example is the windows, now available with a combined  $U$ -value (frame and glazing) of  $0.8 \text{ W/m}^2\text{K}$ .

The aesthetic integration of passive solar, active solar, and photovoltaic (PV) systems is still evolving. Too often, the engineering may be excellent but the resulting appearance not, or vice versa. First semester architectural design principles are also valid for solar systems integration. The resulting “design” should please laymen and not just editors of high-end architecture journals. These concerns were discussed in a session of the Passivhaus conference in Krems AT [19].

Following is a presentation of the three passive solar heating concepts and their variations with example built projects, hopefully which appeal also as “designs.”

## Direct Gain

### Principles



Windows transmit sunlight into the building interior where it is absorbed, and becomes heat. The windows trap the heat in the room and interior construction mass stores some of the heat for the night.

How well the glass transmits solar energy is characterized by its  $g_{\text{value}}$ . A value of 100% would mean all the solar energy gets through the window, i.e., when the window is open. Otherwise, the glass absorbs some of the radiation and is warmed. The warmth is radiated to the ambient and into the room. Since the ambient in winter is colder than the room, it receives more of that heat. Still, some of the heat absorbed in the glass is

radiated into the room. That heat plus solar energy transmitted through the glass comprise the total solar gain. This sum divided by the amount of solar radiation striking the window is the  $g$ -value. Multiple pane glazing systems with selective coating drastically reduce heat loss, but also let less solar radiation into the room. However, the benefit of the lesser heat loss outweighs the reduced solar transmission.

The usefulness of passive solar gains depends strongly on the match or mismatch of solar intensity, occupancy heat gains, and heating demand over the course of a day. Table 1 summarizes characteristics of different window orientations.

Advantages and disadvantages of direct gain:

+	Simplicity	Window construction is highly developed with a long history of passive solar heating experience.
+	Efficiency	Mid-winter solar usability can approach 100%.
+	Economy	People need daylight; buildings need windows, so only marginal cost for better, larger windows must be amortized by energy savings.
+	Aesthetics	Light and warmth from the sun are assets. Fenestration strongly defines the “personality” of a building, hopefully linked to functionality.
–	Overheating	Risk greater than a windowless, mechanically cooled, and ventilated insulated cube. This risk, however, can be calculated and minimized and such a cube is no alternative for providing living quality.
–	Glare	Sunlight on a work surface, computer screen, or poster from Klimt is highly detrimental. Variable, occupant-adjusted shading is essential.

### Components

*Glazing:* Table 2 compares daylight transmission ( $t$ -value), solar transmission ( $g$ -value), and heat loss ( $U$ -values) for a sample of glazings [20]. Exact values are readily available from glass manufacturers’ catalogs. The first three glass types are seldom used today and serve here a reference for comparing modern glass.

**Passive Solar Heating in Built Environment. Table 1**  
Window orientations and characteristics

Orientation	Characteristic
South	Maximum usable winter solar gains. Easiest summer shading
West	Poor solar usability (solar gains follow all day occupancy gains). Overheating risk in summer. Shading more difficult (adjustable vertical elements)
East	Limited solar gains in winter, especially by morning fog. Good solar usability (solar gains after night set-back). Less overheating risk (no direct sunlight after mid-morning)
North	Least solar gains. Greater heat loss (colder microclimate of north side of bld.). Best, daylight orientation, least glare problem. Ideal for offices, school class rooms. Good insulation glass required for comfort near windows
Tilted	Construction complicated, expensive. More difficult to keep weather and water-tight. Greater summer overheating risk (except tilted north). Mounting movable shading elements more difficult
Roof	Maximal daylight by overcast skies. Highest overheating risk (max. solar gains in summer). Difficult to shade. Greatest heat losses in winter accentuated by clear night sky radiation, minimal solar gains on flat roofs

**Passive Solar Heating in Built Environment. Table 2**  
Glass properties

Glass type	t-Value%	g-Value%	U-Value W/m <sup>2</sup> K
Single 3 mm	90	85	5.8
Double	82	75	2.9
Triple	73	65	2.2
Double, low e, Argon	80	60	1.1
Triple, low e, noble gas	76	56	0.6
Double, low e, vacuum	68	50	1.2

*Vacuum glazing:* In multiple pane glazings, heat is transported by radiation between the panes and convection of the gas in the cavity between the glass panes. To improve performance, coatings are applied to the cavity side of the glass. The coating selectively lets more solar radiation through than heat back out. To reduce the convection heat loss, a noble gas, like Argon or Krypton, can be used. Their higher viscosity slows the convection loop. If there is no gas in the cavity, there can be no convection heat transfer. The only problems are to keep the atmospheric pressure from collapse, the glass panes together, and to maintain the vacuum. Small plastic pillars spaced evenly across the glazing area can keep the panes separated. Maintaining the vacuum is addressed in several patented edge sealing technologies. An important benefit of vacuum glazing is its slimness, with a total thickness of 6.5–11 mm depending on the needed glass strength. The gap for the vacuum is only about 0.25 mm. A vacuum between 4 and 10 Torr is used (a pressure unit equal to 1/760 of an atmosphere). This is a relatively weak vacuum; a thermos bottle has 6–10 Torr [21].

*Glazing spacers* in multiple-pane glazing are thermal bridges. Earlier insulating glazing used aluminum spacers. Unfortunately, aluminum is a good heat conductor, so edge losses were high. A next generation used stainless steel spacers, with a lower conductivity. Modern insulating glass units use spacers with a plastic thermal break. The improvement is substantial. Aluminum spacers have linear heat loss ( $\Psi$ ) of 0.07–0.8 W/mK. The  $\Psi$  of a thermally separated spacer (i.e., stainless steel separated with plastic) can be as low as 0.04 W/mK. Table 3 illustrates how strongly the linear thermal bridging of the edge spacer affects the overall U-value of the glazing, depending on glass area [22].

Assumptions	$U_{\text{frame}} = 1.6 \text{ W}/(\text{m}^2\text{K})$
	$U_{\text{glass}} = 1.1 \text{ W}/(\text{m}^2\text{K})$
	$\Psi = 0.070 \text{ W}/(\text{mK})$

*Window frames* are the weak thermal component of windows with highly insulating glass. The frame has a worse U-value and of course it blocks the sunlight. So, frames with a small profile are desirable. Frames with some form of thermal break to interrupt the heat path

**Passive Solar Heating in Built Environment. Table 3**  $U_{\text{window}}$  value of different windows sizes, including the effect of the edge spacer

$w \times h$ (mm $\times$ mm)	$A_{\text{window}}$ (m <sup>2</sup> )	Perimeter (m)	$A_{\text{window}}/\text{Perimeter}$	$U_w$ (W/(m <sup>2</sup> K))
400 $\times$ 800	0.32	2,400	0.133	1.8
1,300 $\times$ 1,300	1.69	5,200	0.023	1.5
1,230 $\times$ 1,480	1.82	5,420	0.024	1.4
2,750 $\times$ 2,500	6.88	10,500	0.014	1.3

are desirable. Even the  $U$ -value of a solid wooden window must today be judged as optimal for very low energy buildings, as can be compared in Table 4. Note that, with the exception of the aluminum frames, good  $U$ -values can be obtained for all materials. Exact  $U$ -values should be obtained from manufacturers because the values given here can vary relative to specific products. Also, of course, insulation value is only one selection criteria among many, i.e., strength to resist wind forces, life span, and maintenance costs.

*Fixed shading* by roof overhangs is promoted as a solution for south facades. This must be questioned for climates with overcast winters. By an overcast sky, the most daylight comes from the zenith. Therefore, fixed overhangs block daylight during long gray periods when daylight is most desired. For such climates, moveable shading is superior.

To estimate the adequacy of a south-facing overhang, the highest and lowest noon sun angles (21st June and December) are calculated as follows:

21 June	$90^\circ - \text{latitude} + 23.45^\circ$
21 December	$90^\circ - \text{latitude} - 23.45^\circ$

Taking Zurich (latitude approx.  $47^\circ\text{N}$ ) as an example, the highest and lowest sun angles are  $66.45^\circ$  and  $19.55^\circ$ , respectively.

While this is a good first estimation for designing a shading geometry, the problem is that the sun has a lower angle before and after solar noon. An overhang should extend horizontally beyond either side of the window to give diagonal shading as the sun rises, falls, and moves laterally before and after noon.

**Passive Solar Heating in Built Environment. Table 4** Example window frame constructions and thermal properties

Frame construction	$U_f$ frame (W/m <sup>2</sup> K)
Solid wood <sup>1</sup>	1.3
Wood-aluminum <sup>1</sup>	1.2
Wood with air cavities <sup>2</sup>	1.1
Plastic <sup>1</sup>	1.1
Aluminum <sup>3</sup>	2.2
Aluminum with break <sup>3</sup>	0.9

<sup>1</sup>EgoKiefer, CH-9450 Altstätten SG, [www.swiss-topwindows.ch](http://www.swiss-topwindows.ch)

<sup>2</sup>Tischlerei Sigg GmbH, AT-6912 Hörbranz, [www.passivhausfenster.at](http://www.passivhausfenster.at)

<sup>3</sup>Schüco/Jansen AG, CH-9463 Oberriet, [www.jansen.com](http://www.jansen.com)

East- and west-facing windows need vertical shading since at sunrise and sunset the sun will get under any overhang. Vertical shading elements that can be rotated away from the lateral movement of the sun are best, to allow shading and some view concurrently.

*Mass* increases the effectiveness of passive solar gains and is especially effective if directly sunlit (primary mass). It is up 150% more effective than secondary mass heated indirectly by the room air [10]. A recommendation for middle European-like climates is to provide 2,800 kg of mass per m<sup>2</sup> of window area [23]. Another recommendation is that for each m<sup>2</sup> of south-facing glass above 7% of the floor area, there should be between 6 and 8 m<sup>2</sup> of exposed thermal mass. An example would be a 200 m<sup>2</sup> house with 20 m<sup>2</sup> of south facing glazing. 6 m<sup>2</sup> of that glazing will require 36–48 m<sup>2</sup> of solar-exposed thermal mass [24].

**Passive Solar Heating in Built Environment. Table 5** Heat storage properties of common construction materials [25]

Material	Density $\rho$ (kg/m <sup>3</sup> )	Conductivity $\lambda$ (W/mK)	Thermal capacity $c$ (Wh/kgK)	Volumetric heat capacity Wh/m <sup>3</sup>
Metamorphic stone	2,800	3.5	0.26	728
Sedimentary stone	2,600	2.3	0.22	572
Clay	1,700	0.9	0.24	408
Sand, gravel	1,800–2,000	0.7	0.22	418
Concrete reinforced	2,400	1.8	0.3	720
Concrete aerated	1,000	0.3–1.0	0.3	300
Interior plaster	1,400	0.7	0.26	364
Gypsum board	900	0.21	0.22	198
Wood (pine, fir)	450–500	0.14	0.55–0.66	287
Wood (oak)	700–800	0.21	0.55–0.66	454
Wood (fiber board)	800	0.17	0.7	560

Also, note for day–night heat storage, thickness greater than approximately 10 cm will not increase the solar usability.

If the primary mass, i.e., a stone floor or brick wall, is a dark color, it will absorb the solar radiation better, but the impact on daylight distribution must be considered. Light-colored sunlit surfaces, especially floors or side walls, are essential to diffuse daylight deeper into a room. Such surfaces should have a mat color to avoid glare.

How well a materials stores heat is indicated by its capacity. Table 5 gives the physical properties of some construction materials to compare their effectiveness as thermal storage [25].

**An Example Building:** A very impressive example of passive solar heating with windows is a single family house built at 900 m above sea level in Trin, CH. It has no auxiliary heating, not even a wood stove. The architect, Andrea G. Ruedi, matched a very large, south-facing window area (46 m<sup>2</sup>), very plentiful mass inside the insulated envelope (Fig. 13). The envelope is wooden frame construction to simplify achieving a high insulation value (0.14 W/m<sup>2</sup>K); the interior incorporates limestone bricks and concrete (see



**Passive Solar Heating in Built Environment. Figure 13** Extreme example of a direct-gain house in Trin/CH with no auxiliary heating (architect Andrea Ruedi, CH-7000 Chur)

Table 6). The room temperatures were measured in a research project over the heating season, with no auxiliary heating. They varied between 18 and 23°C. Only very rarely did the temperature fall below 19°C. The theoretical annual heating energy, were 20°C maintained, was calculated to be less than 30 L of heating oil equivalent (1.1 kWh/m<sup>2</sup>a) [26].

**Passive Solar Heating in Built Environment. Table 6**  
Properties of the solar House in Trin CH

Properties	Trin solar house	A reference house
South window area	48 m <sup>2</sup> , (8% frame)	28 m <sup>2</sup> , (20% frame)
South-facing window to façade proportion	56%	33%
Interior storage mass	277 t	190 t
Wall construction	Exterior: insulated wooden frame, Interior: limestone masonry	15 cm limestone 30 cm cellulose insulation

### Design Advice

#### Maximize solar gains

- Orient direct-gain window gains between  $\pm 45^\circ$  from south
- South window to façade ratio: 30–50%, not more
- Large, uninterrupted glass areas (to minimize frame and glass edge losses)
- Window  $U_{\text{window}} < 1.0 \text{ W/m}^2\text{K}$  (including frame) and a good  $g$ -value ( $> 50\%$ )
- Account for winter sun blockage by neighboring buildings, trees, terrain

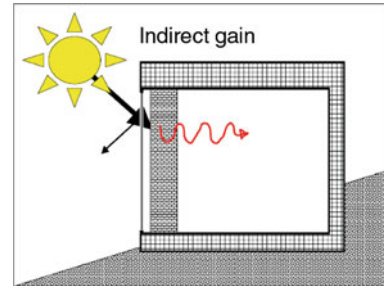
#### Maximize usefulness of passive solar gains

- Interior construction with adequate sunlit (primary) mass
- Room interior finishes light color to maximize light distribution
- Open floor plan. Largest rooms on south-side (small rooms overheat faster)
- Auxiliary heat control responsive to passive solar gains
- Shading elements: horizontal for south, vertical for east/west
- Adjustable shading to allow concurrent view and ventilation
- Exterior sun shading to keep absorbed heat outside

- Generous operable window area with max. height difference to induce natural ventilation (diurnal cooling where possible)

### Indirect Gain

#### Principles



Sun warms building walls and roofs but normally the heat is radiated and convected back to the ambient. By protecting the surface behind glass, heat can be trapped within the construction, stored or transported into the building to reduce auxiliary heating demand.

Several variations of this concept have been built including: the mass wall, mass roofs, transparent insulation, and solar insulation. Of the many innovative concepts, only a few have survived into the present, but with rising energy prices and availability of new high-performance components, these concepts can be promising.

How much passive solar heat gains can reduce purchased heat demand depends on the intensity and timing of the sunlight, occupancy heat gains, and room temperatures desired. Table 7 summarizes characteristics of different window orientations.

#### Advantages and Disadvantages:

+	Simplicity	The concept is simple, some variations do not have any moving parts.
+	Aesthetics	Most systems include large glass areas, which can be integrated with window areas into an attractive transparent façade concept.
–	Inefficiency	Heat losses from the solar energy captured in the collector are radiated both to the inside and the ambient, reducing system efficiency.

+	Natural cooling	The chimney effect of an indirect gain system can draw cooler air from a ground channel or the north facade through the building for summer comfort.
–	Complexity	Many systems need seasonal shading to avoid summer overheating. Some require dampers to regulate and direct air flows.
–	Cost-benefit	Complex systems proved, in most cases, to be too expensive for the energy benefit.

Following are four variations of indirect gain systems, mass walls, mass roofs, transparent insulation, and solar insulation.

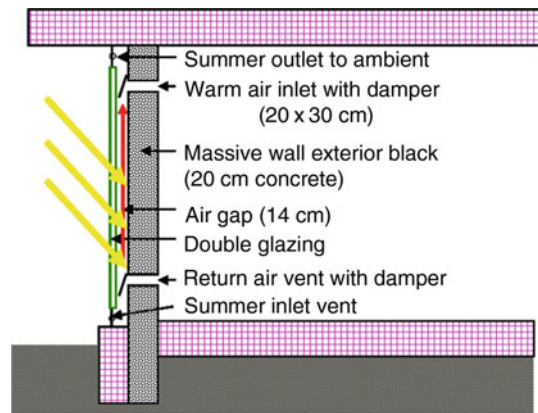
**Mass Walls**

A stone, concrete, brick, or adobe wall will absorb and store solar heat, but the heat is rapidly radiated and convected back to the ambient with little or no benefit to the heated building space. If the wall is protected behind glass, the heat is better retained. With a time delay, much of the heat can then penetrate and be radiated and convected to the room behind the wall (Fig. 14). This is the principle of the solar wall, patented in 1881 by Edward Morse, an American Botanist. In 1964, a French engineer, Felix Trombe and architect Jacques Michel built such a wall to demonstrate this principle. Since then, the mass wall or Michell–Trombe Wall has become popularly known as the Trombe Wall [27].

The wall has been built in two variations. In the unvented version, the wall delivers heat to the room only by conduction and then radiation from the wall surface, with up to an 8–10 h time delay, depending on how massive the wall is. In the vented version (Fig. 14), the vents open when the air in the cavity is sun-warmed. The air circulates into the room at the top of the wall and returns to the cavity through slots at the bottom. This variation delivers heat sooner so is better for east-facing walls, and would not be good for a west-facing wall. To prevent back circulation of cold air in the gap into the room at night, dampers are needed. One solution was a Mylar film damper, which simply flapped open or was pressed closed against a wire mesh

**Passive Solar Heating in Built Environment. Table 7** Indirect gain system orientations and characteristics

Orientation	Characteristic
South	Maximum usable winter solar gains. Fixed overhang possible
West	Less solar gains compared to south-facing facades. Heat delivered to space at time of day when least needed, so storage important. Greatest risk of summer discomfort
East	Limited solar gains compared to south-facing facade. Less overheating risk (no direct sunlight after mid-morning). If storage included, heat delivered at mid-day when least needed
North	Least solar gains, questionable cost-benefit
Roof	Maximal night-sky cooling in summer, least benefit in winter. Steeper roofs intercept winter sun better



**Passive Solar Heating in Built Environment. Figure 14** Concept of the Michel–Trombe mass wall

by the air pressure. In summer, dampers could be opened at the top and bottom of the air gap to vent it to the outside. Alternatively, only the top damper could be opened and a north window of the house opened. The chimney effect of the mass wall draws cooler air from the north side of the house, across the room and exhausts it out the top of the mass wall to the ambient.



**Passive Solar Heating in Built Environment. Figure 15**  
The prototype Michel-Trombe wall house in Odeillo, FR  
(source: Robert.hastings@aeu.ch)

The first project in Odeillo France (Fig. 15) was subsequently copied at Los Alamos, NM, United States, where it was instrumented and a computer model of its physics calibrated. Versions were then built in the 1980s in middle Europe. The performance during long, overcast winter periods was disappointing, while in summer the rooms behind the wall were too warm. Another solution was needed for this climate.

### Mass Roof

An innovative alternative to a solar mass wall is a solar mass roof. It can provide both winter heating and summer cooling. One innovative concept uses a series of roof water bags (like a water bed) and moveable insulation panels. During the winter days, the insulation is slid back on its tracks, so that the water is sun-warmed. Nights, the insulation is rolled back over the water bags, which then conduct heat through the steel deck ceiling to be radiated down to the rooms beneath. In summer, the process is reversed. Nights the cover is removed and the water is cooled by radiation to the sky, days the insulation is slid in place and the rooms below are cooled by the cold water bags.

A prototype house was built in 1973 by the inventor of the concept, Harold R. Hay (Fig. 16). The three-bedroom, two-bath structure in Atascadero California was constructed and monitored with funding from the US department of HUD. It was the first documented



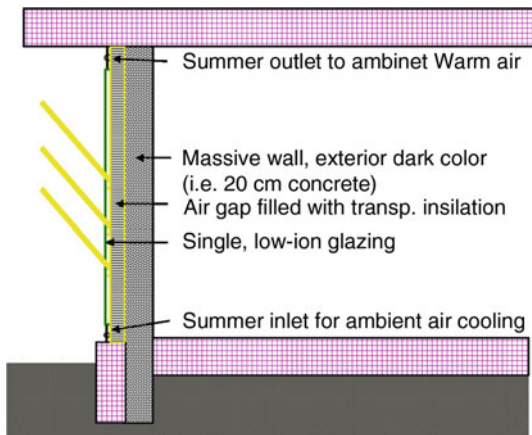
**Passive Solar Heating in Built Environment. Figure 16**  
A mass roof: the Skytherm house in Atascadero California  
by Harold Hay (photo source: Evelyn and Harold Hay Fund  
at Cal Poly, San Luis Obispo, USA)

100% passive solar heated and cooled building and the only instrumented passive solar house in operation during the 1973 energy crisis. To engineer the system, the then new generation of computer simulation tools was used (simulations for this project done by Phil Niles) [28]. It could be worthwhile to reexamine this concept, given the materials and insulation systems available today.

### Transparent Insulation

The transparent insulation wall (TWD) improved the performance of the mass wall concept by addressing one of its weaknesses. In the cavity, the air warmed by the black surface of the solar wall rises, while cooler air against the surface of the glass falls. The resulting circulation loop transports heat from the wall to the glass where it is then lost by conduction to the ambient. In the transparent insulation system, the air gap is filled with some form of transparent cellular structure, inhibiting the convective loop (Fig. 17). The infill material is typically a cylindrical, rectangular, or honeycomb geometry, which directs the light in multiple reflections to the mass wall at the back. A small vertical gap between the TWD and glass should be maintained, to allow moisture to diffuse and prevent the TWD from being in direct contact with the hot absorber surface.





**Passive Solar Heating in Built Environment. Figure 17**  
Transparent insulated wall concept

In non-heating months, vents to the ambient can be opened at the top and bottom to the ambient to cool the wall. These proved to be difficult to keep air tight in winter and added cost. Some form of shading for the wall was needed. Window roller blinds are effective, but make the whole system prohibitively expensive. An innovative solution was to use fixed, metal micro louvers with the fins set at an angle to block high summer sun angles. They, unfortunately, also reduced winter solar performance and were expensive. A third variation uses fused transparent spheres 2–3 mm in diameter as the glazing, applied like transparent stucco. These let less solar energy through, but with the benefit of much better summer comfort behind the wall. The area of the glazing patches is for physical reasons limited [29].

Several transparent materials and geometries have been used to fill the air gap, including extruded PMMA-Capillaries or Polymethylmethacrylat (Plexiglas), extruded PC Polycarbonate (Makrolon) capillaries or honeycomb forms, and extruded polycarbonate multi-cell panels. Critical for the selection are the upper temperature tolerance and UV-stability. Some materials are stable up to 120°C, other to only 90°C. A good overview of products and properties is available from the association of TWD manufacturers [29].

A well-publicized example of a TWD-building under extreme conditions is a Swiss alpine hostel at



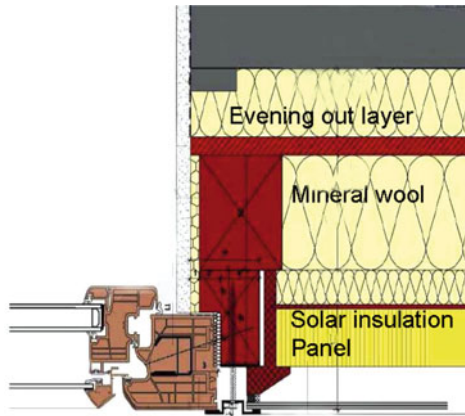
**Passive Solar Heating in Built Environment. Figure 18**  
Swiss TWD house at Hundwiler Höhe (architect and photo source: P. Dransfeld, [www.dransfeld.ch](http://www.dransfeld.ch))

Hundwiler Höhe at an elevation of 1'306 m above sea level (Fig. 18). It was built in 1995 and 42 m<sup>2</sup> of prefab TWD-Modules 130 × 90 × 18.5 cm (h × w × d) were used. No summer solar protection was needed at this altitude. The 185 cm thick TWD wall construction (outside to inside) is as follows: 8 mm framing projection beyond the glass, 4 mm glass, 30 mm gap, 120 mm transparent insulation, 8 mm absorber, 15 mm air gap between the absorber and wall. This air gap was needed to provide the needed tolerance for mounting the prefab TWD modules. The gap was estimated to cause a 10% reduction in efficiency, which was considered acceptable. Simulations indicated that the temperature in the TWD construction should not exceed 80–90°C, well within the 110°C tolerance of the TWD material used [30].

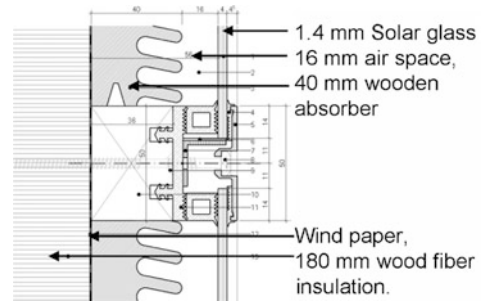
### Solar Insulation

The solar wall concept is the simplest of the indirect gain systems and perhaps, therefore, most economical.

The goal of this concept is to provide dynamic wall insulation. During the day, air chambers in the cavity protected by glass are warmed by the sun. Nights, the cavity slowly cools down. The air chambers together with the glazing to the outside and insulation to the inside help reduce heat loss from the building.



**Passive Solar Heating in Built Environment. Figure 19**  
Wall section of a cellulose solar wall insulation system  
(Redrawn based on a figure in the report:  
Domenig-Meisinger et al. [31])



**Passive Solar Heating in Built Environment. Figure 21**  
A solar insulated wall detail of the routed wooden Lucido system (source: Lucido Solar AG Solares Bauen, [www.lucido-solar.com](http://www.lucido-solar.com))

The cellulose system (GAP), shown in Fig. 19, achieves on south facades a dynamic  $U$ -value of  $0.08 \text{ W/mK}$  in middle Europe [31].

A well-publicized example using this concept is an Austrian Apartment building on Makartstraße, Linz (Fig. 20). To minimize disturbing the tenants, prefabricated wall panels including the solar walls, windows, sun-shading systems, and ventilation channels were mounted. The south facades achieve a dynamic  $U$ -value of  $0.08 \text{ W/mK}$  averaged over the heating season. As a result of a combination of this wall system and other measures, the heating demand could be reduced by 92% to  $13.4 \text{ kWh/m}^2\text{K}$  [31] and [32].

The routed wooden system (Lucido) entraps air in inward-sloping slots (Fig. 21). Important in this and also the cellulose board system is that the wall behind the solar wall be well insulated. A benefit of the wooden absorber is that, being weather protected, it can be left natural and hence conveys the character of a wooden façade.

### Design Advice

Following is design advice for temperate climates. In mild climates, these systems might make it possible to reduce auxiliary heating demand to zero, but summer comfort strategies must be well done. In a humid, hot climate, this system make no sense, nor is performance likely to be good in northern, very cold and weak-sun winter climates. Many



**Passive Solar Heating in Built Environment. Figure 20**  
Apartment building with the GAP solar insulation on  
Makartstraße, Linz AT (photo source: S. Grünewald and  
S. Rottensteiner)

Two construction variations exist for creating the insulating air chambers: a type of treated, corrugated cardboard and wood routed with horizontal slits.

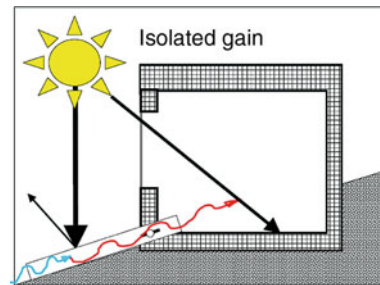
projects were built in temperate climates but a market breakthrough has not yet occurred. The energy they save for the investment is high compared to energy from cheap fossil fuel.

- Because indirect solar gain is less efficient than direct gain, a large collection area is needed. South-facing orientations are most sensible. Depending on the desired timing of heat release, east- or west-facing solar mass walls are possible, but the absolute amount of delivered heat will be smaller.
- Overheating is a risk in mass wall systems, so summer sun shading and venting are important. The solar insulation concept has the comfort advantage of having an insulated wall separating it from the building interior.
- Durability was a problem for early prototypes, including untight vent dampers and degradation of sun-exposed wooden framing. A typical greenhouse construction with a metal cap to protect exterior wood is one solution. Transparent insulation can deform at high temperatures, so the right material must be chosen for the design, or reliable shading provided. Freeze protection UV-durable materials are obvious requirement for the roof-mass system using water.
- The thickness and density of a mass wall and hourly solar radiation should be calculated to dimension a mass wall to deliver its heat to the room when desired.
- System performance might be improved by very good insulating glass. This is a trade-off of  $g$ -value and  $U$ -value in the context of the economics. Single glazing in low-iron glass could still be the best solution (maximizing the  $g$ -value).
- The mass roof system is only plausible in clear-sky climates with both a heating and cooling demand.
- The room side of the mass wall or transparent insulation wall should not be blocked by furniture. For the solar insulation system, this is not an issue. The room surface of all the wall concepts can be any color desired.
- Prefabrication can provide cost savings for a second project, not necessarily the first project. The benefits are shorter on-site erection time, less disturbance of occupants, and better quality control, which can lead to better durability.

- Indirect gain systems are well suited for building renovations.
- These systems have gone through development pains; valuable experience is available from the project designers, research institutes, and the manufacturers of system solutions. Homework to assure a new project starts from the state-of-the-art is essential!

**Isolated Gain/Hybrid**

**Principles**



Solar energy is collected outside the insulated envelope of the building, then transported as heat by convection into the building or into storage. Two variations are considered here: solar air systems (with or without mass) and sunspaces.

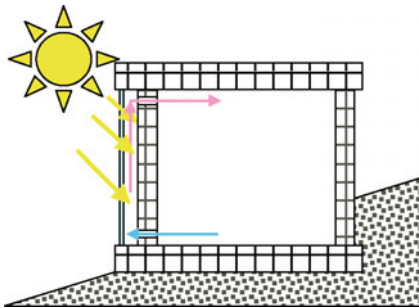
The orientation of a sunspace, like any room, depends on view and when sunlight is wanted. For solar air systems, design issues are similar to those of indirect gain systems. Buildings uninhabited and kept at a minimum temperature much of the year (i.e., vacation homes) are an ideal application.

Advantages and disadvantages:

+	Simplicity	These systems are simple and reliable.
+	Dependability	The gravity-driven solar air systems work without moving parts. However, a small fan would improve the efficiency and can easily be PV-powered, making the system immune to grid power interruptions.
+	Economy	Reduced purchased energy costs and reduced wear from less running time of the auxiliary heating system, extending its lifetime.

+	Function	Sunspaces are built for the space they provide, energy savings are only a fringe benefit, so in effect a bonus.
+	Overheating	Isolated solar gains systems, because they are outside the insulated building, are advantageous regarding summer comfort. Sunspaces need large, low, and high ventilation openings, and effective shading and glare protection.
+	Natural cooling	The chimney effect of an isolated solar collector or sunspace can draw cooler air from a ground channel or the north facade through the building.

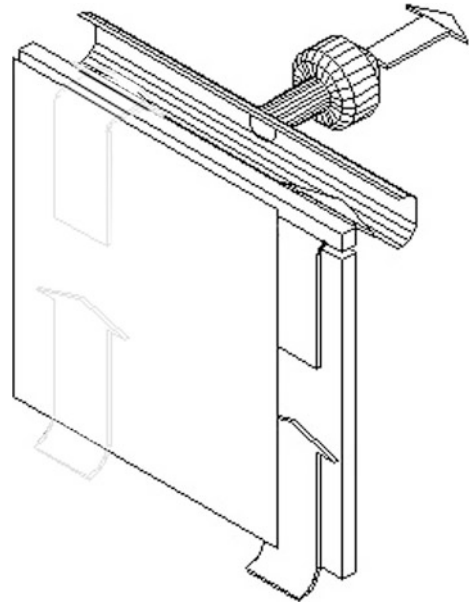
### Solar Air Systems



A solar air system is in effect a sunspace with the depth reduced to a few centimeters. The principle is the same as an active solar water system, except that heat is transported from the collector to the point of need or storage by air. Two variations are reviewed here: systems that are directly coupled to the building and systems in which the sun-warmed air passes through mass before entering the building. All together, six system types were researched in a project of the International Energy Agency. Out of this work, a design handbook for solar air systems [33] and a book of example built projects [34] were published. The other four system types typically require an electrical fan, dampers, and more complex control systems to function, so are not included here under passive systems.

### No-Mass Solar Air Systems

These systems operate on the principle that sun-warmed air in a vertical or upward sloped volume behind glass will rise. This warm air can then be



Passive Solar Heating in Built Environment. Figure 22  
Diagram of a free convecting façade solar air collector

channeled through the insulated wall of a building to provide solar heating (Fig. 22). Two variations are possible for the air supply at the bottom of the collector:

- If the opening is to the ambient, the collector can deliver sun-warmed fresh air to the building.
- If the opening is to the building, the collector delivers recirculated, higher temperature air than the first variation. In either configuration, in summer an outlet at the top can be opened to the ambient to exhaust the hot air.

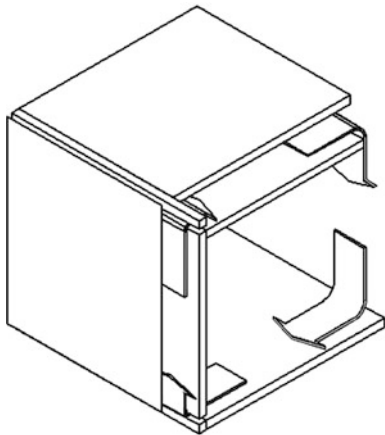
**Example:** Figure 23 shows a solar air system to keep a vacation home in Koroni GR heated to a low level, ventilated and dry during periods of vacancy. Two collectors, each 6 m<sup>2</sup>, circulate up to 200 m<sup>3</sup>/h of fresh air into the house. A small PV panel (50 wp) integrated into a corner of the air collector powers a small fan to increase the efficiency of the system, which has been in operation since 2004.

### Mass Solar Air Systems

This concept is like the no-mass solar air system described above, except that the solar-heated air



**Passive Solar Heating in Built Environment. Figure 23**  
A solar air heater for a vacation home on a Greek island  
(photo and system information: [www.grammer-solar-bau.de](http://www.grammer-solar-bau.de))



**Passive Solar Heating in Built Environment. Figure 24**  
A solar air collector linked to air channels in the building structure

circulates through the building structural mass before entering the room (Fig. 24). In this way, the air enters the room at not as high a temperature, and the mass continues to radiate the stored heat after sunset. The mass may be a concrete ceiling or floor (hypocaust) or walls (murocaust) with air channels. The system can function with only free convection of air movement [33].



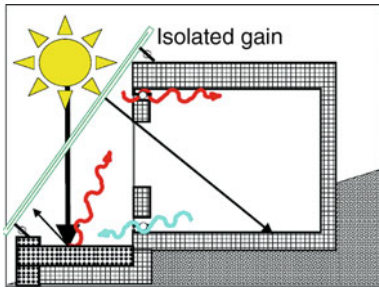
**Passive Solar Heating in Built Environment. Figure 25**  
Solar apartment buildings in Marostica by Barra Costantini  
(photo: Gianni Scudo)

**Example:** Figure 25 shows an apartment building in Marostica, Italy (20 km from Vicenza) with a façade integrated passive solar air system. The sun-warmed air in the collector rises naturally and circulates through channels in the concrete ceiling/floor structure before entering the apartments in the north-facing rooms. The concept was developed by Barra-Costantini. Each 84 m<sup>2</sup> apartment is heated by 16 m<sup>2</sup> of collector. Each m<sup>2</sup> of collector is estimated to contribute about 100 kWh/a [34].

#### Design Advice

- The collector can be mounted below floor level, i.e., in the case of a building with an above-grade basement. The height difference strengthens the free convection.
- No-mass solar air systems are well suited for buildings often vacant, which need to be tempered and supplied fresh air. In permanently occupied buildings, mass is essential to maximize the usefulness of the collector gains and avoid overheating.
- Dampers are essential to prevent reverse-flow and cooling of room air into the collector at night.
- A small fan can increase system efficiency. Commercial solar air systems with PV-powered fans are available.
- The collector and solar-heated air channeling require good engineering. Consult the literature to not have to reinvent the wheel [33, 34].

## Sunspaces



Sunspaces became popular element of passive solar architecture. They were designed as an architectural feature which, in addition, reduced purchased energy consumption in several ways:

- Passively by creating a warm buffer zone on the south side of the house, reducing wall and window heat losses.
- Passively by occupants simply opening house windows and doors into the sunspace when its temperature exceeded the house temperature. Alternatively, a small thermostat could open a damper and switch on a fan to automate this.
- Actively, when sunspace supplied sun-warmed air to a mechanically ventilated building. Alternatively, the sunspace air could be ducted to a heat exchanger to warm incoming ventilation air for the building.

Today, in highly insulated buildings, a sunspace's buffering effect is no longer a significant energy saving. In mechanically ventilated buildings with heat recovery, the benefit of heating the incoming air is also less significant, but still a benefit. Sunspace heated air can exceed room temperature, thus supplying useful heat. A sunspace can increase purchased energy consumption if occupants heat it to near room temperatures. Comfort expectations of a sunspace must be less than for rooms.

**Example:** The Wydacker row houses in Zollikofen (Bern) CH are earth sheltered to the north and protected behind a sunspace to the south (Fig. 26). This construction provides energy benefits and protection from nearby street noise.

Each house has a 108 m<sup>2</sup> sunspace with a 57 m<sup>2</sup> of insulated glass ( $U=2.9 \text{ W/m}^2\text{K}$ ) at a 60° slope oriented 20° west of south. Being slightly west of south is



**Passive Solar Heating in Built Environment. Figure 26** Attached sunspaces as part of the concept of low energy row housing in Bern (Architects, AARPLAN; Bern, CH)

beneficial. Frequent morning fog reduces solar radiation mornings compared to afternoons. Sun shading is provided by a roller shade beneath the glazing. The concrete block wall of the house and concrete pavers over gravel provide thermal mass for the sunspace. The measured heating energy consumption of the houses was 37 kWh/m<sup>2</sup>a, which for the year 1995 was excellent performance [35].

### Design Advice

- Insulating glazing for both the sunspace (minimize freeze risk for plants) Insulating glazing for the house to minimize heat loss to the sunspace.
- Sunspace frame out of laminated wood to be dimensionally stable and metal exterior cap to reduce weathering. Alternatively, aluminum framing with a thermal break.
- Two or more story sunspaces offer d more collection area for the enclosed volume. Comfort is better because stack effect ventilation improves with stack height).
- Large operable sash at base of the sunspace and at its top, ideally with rain sensor-activated closers. Rule of thumb: Minimum 1/6 glass area operable.
- Sun shading on exterior most effective, on interior it is less subject to wind damage and weathering. Interior sunshade installed min. 10 cm below glass

so gap acts as thermal chimney between operable low inlet and high outlet sashes.

- As in direct gain systems, thermal mass helps reduce temperature swings (i.e., minimize hours below freezing).
- A freeze-protecting heater with thermostat activation when the sunspace temperature falls below 4°C can help protect plants. If the sunspace is designed, the purchased energy for this is well worth the plants.

### Future Directions

The future directions to using passive solar energy can best be forecast by reviewing technical and political events in the past. In the early twentieth century, when production of glass in large formats became possible, architects began experimenting with large glazed areas. The resulting buildings were uncomfortable to occupy in winter and expensive to heat. First, with the introduction of insulating glass, a net passive solar heat gain in winter became possible. After World War II, oil became plentiful and cheap and interest in solar dwindled. Then, the oil crisis of 1973 renewed interest in finding alternatives to fossil fuels. New solar building concepts evolved under the collective term “passive solar heating.” The first of annual “National Passive Solar Conference” was held in Sante Fe, NM (United States) A US federal department (HUD) held a landmark national competition for innovative passive and active solar building concepts. Interest and built projects spread from the sunny southwestern United States across the entire continent. By the next decade, passive solar heating concepts were being applied by architects in Europe as well. Passive solar concepts, originating from inventive individuals, became a topic for national research institutions. Test cells and buildings were monitored, computer models developed, and engineering handbooks written.

Today, the term, “passive solar buildings” is less commonly heard. The similar sounding concept “Passive House” now enjoys international attention. The new, future-oriented trend is sustainable buildings [36, 37], net-zero-energy buildings, carbon neutral buildings, and even energy-plus buildings. Well-designed new buildings constructed to such high standards need very little heat, and so passive solar gains are less beneficial than before, but still an asset. Such solar

gains help delay when the heating season finally begins and end the heating season earlier. Passive solar gains are also a major heat source when a building is unoccupied during the day or for extended periods. The daylighting aspect of direct solar gain concepts will continue in the future, being a major appealing factor.

Future buildings must address new requirements. There will be

- More elderly people (demographics) with greater comfort expectations.
- High energy prices, regardless whether the source is increasingly scarce fossil fuels, electricity, or renewable energy.
- Less disposable income because salary increases will not match the inflation of energy costs affecting prices of all goods and services.

New technological developments will offer new possibilities for meeting these requirements. Likely developments may include:

- Nanotechnology selective coatings for glazings to allow larger passive solar collection areas in winter and no overheating in summer. Similarly, material science will deliver high-performance coatings for absorber surfaces.
- Vacuum technologies for both window glazing and as very compact insulation for indirect or isolated gain systems.
- Intelligent, self-learning control systems for switchable property components, responding to solar intensity, ambient temperatures, and programmable occupant comfort profiles.
- Chemical thermal storage for heat or “cold” to provide very compact and high density, compact storage with no losses during storage.
- Building skins, which produce both electricity and low-temperature heat amplified by high-efficiency heat pumps for space and water heating.

Finally, in the future as in the past, political developments may result in supply interruptions. In any event, as the world-known reserves diminish, prices will not steadily and gradually increase. Large price swings amplified by speculation can be expected.

These changing occupant requirements, technical developments, and possible political and oil market events will change how new buildings are constructed

and existing buildings renovated. They will more effectively draw energy from the environment, providing heat, light, and “cool” with less external energy input. Buildings and climate will no longer be combatants, but allies working together. The needed investment and maintenance costs will have to be low, simply because there will be less disposable income. Passive solar heating was, is, and will be an important strategy for achieving low-energy buildings offering excellent living quality.

### Acknowledgments

The author thanks in particular three institutions in his career, whose support made possible the personal experiences to write this section.

The Swiss Federal Office of Energy, Buildings Program (in particular, Gerhard Schriber).

The International Energy Agency, Solar Heating and Cooling Program and, in particular, all the researchers and architects who, with such dedication, worked together in research tasks over the decades and the founder of the Program, Fred Morse.

The Donau University-Krems, Department of Buildings and Environment (in particular: Peter Holzer, who convinced me to become a professor so people would believe my stories).

### Bibliography

1. Feist W (2007) Certification as “Quality Approved Passive House” criteria for residential-use passive houses. [www.passiv.de/07\\_eng/php/Criteria\\_Residential-Use.pdf](http://www.passiv.de/07_eng/php/Criteria_Residential-Use.pdf)
2. Hastings R (ed) (1994) Passive solar commercial and institutional buildings – a sourcebook of examples and design insights. Earthscan, London. [www.earthscan.co.uk](http://www.earthscan.co.uk). ISBN 0-471-93943-9
3. Broadhurst T (ed) (2010) Saltbox, 1650–1830. <http://www.oldhouseweb.com/forums/viewtopic.php?f=4&t=26402&p=234896&hilit=Saltbox#p234918>
4. Simon MJ (1947) Your solar house. Simon and Schuster, New York in cooperation with Libbey-Owens-Ford Glass Co., (Pilkington, North America, Inc.)
5. Butti K, Perlin J (1980) A golden thread: 2500 years of solar architecture and technology. Cheshire Books, Palo Alto. ISBN 0-917352-07-6
6. Franklin Research Center (1979) The first passive solar home awards. US Dept. of Housing and Urban Development, Office of Policy Development and Research, available from US Government Printing Office, Washington, DC
7. ASES (1977) 1st National passive solar conference. American Solar Energy Society, Boulder. [www.ases.org](http://www.ases.org)
8. Balcomb D, Jones R (1980) Passive solar design handbook: vol 1 – passive solar design concepts. Total Environment Action, Inc.; vol 2 – passive solar design analysis. Los Alamos Scientific Laboratory; vol 3 – passive solar design analysis. Los Alamos National Laboratory. Available from National Technical Information Service, Alexandria
9. Hastings SR (1977) Three proposed typical house designs for energy conservation research, NBSIR 77-1309. National Institute for Standards and Technology (NIST), Gaithersburg
10. Mazria E (1979) The passive solar energy book, Expanded professional edn. Rodale Press, Emmaus. ISBN 0-87857-238-4
11. Loftness V (1978) Regional guidelines for building passive energy conserving homes. US Department of Housing and Urban Development, Washington, DC, HUD-PDR-355
12. Hastings R, Crenshaw R (1977) NBS BSS 104 window design strategies. National Technical Information Service, Alexandria
13. Porteous Colin, MacGrego Kerr (2005) Solar architecture in cool climates. Earthscan, London. ISBN 10: 1-84407-281-1 and ISBN-13: 978-1-84407-281-1
14. Hestnes AG, Hastings R, Saxhof B (eds) (1997/2003) Solar energy houses – strategies, technologies and examples. Earthscan, London. ISBN 1-902916-43-3
15. Larsen S (2010) Built examples: house Frei, Architekturbüro Larsen, Erlachstrasse 25, AT-6912 Hörbranz. <http://www.solarsen.com>
16. SIA (1989) Demonstrationsprojekt Schulhaus Gumpenwiesen, Reihe “Planungsunterlagen zu Energie und Gebäude”, SIA-Dokumentation; D035. SIA Schweizerischer Ingenieur- und Architekten-Verein, Zürich
17. Hastings R (ed) (2010) Lessons from exemplary housing renovations, IEA SHC 37B. [www.iea-shc.org/task37/](http://www.iea-shc.org/task37/)
18. METEOTEST (2010) Meteoronorm 6.1. Fabrikstrasse 14, CH-3012 Bern. [www.meteoronorm.ch](http://www.meteoronorm.ch)
19. Hastings R (2004) Passivhäuser und Lebensfreude. In: Eighth European Passivhaus Tagung und Messe, Krems, 16–17 Apr 2004. [www.passivhaustagung.at](http://www.passivhaustagung.at)
20. Troska Christoph (2009) Vom einfachglas zum vakuumisoliertglas. Pilkington, Wikon
21. Rattner E (2008) Revolutionary vacuum glass. <http://thefutureofthings.com/news/1167/revolutionary-vacuum-glass.html>
22. Hermes M (2002) Aktuelles aus dem Regelwerk. In: Die neue wärmedämmtechnische Bewertung von Fenstern mit Einführung der EnEV 2002, Teil 1: Grundlagen. Integratio® – Forum. [www.fensterberater.de/PagEnEV1.htm](http://www.fensterberater.de/PagEnEV1.htm)
23. ZIEGE (2010) Wärmespeicherung – planungsregeln. Verband Österreichischer Ziegelwerke, Wien. [www.ziegel.at](http://www.ziegel.at)
24. Chiras D (2002) The solar house: passive heating and cooling. Chelsea Green, White River Junction. ISBN 1931498121
25. SIA (1981) SIA 381/1 baustoff-kennwerte. Schweizerischer Ingenieur- und Architektenverein, Zürich also copyright 2000 SIA
26. Hässig W, Hardegger P (1996) Messprojekt direktgewinnhaus trin. Forschungsstelle für Solararchitektur, Zürich, Dokumente: 194637.pdf. [www.bfe.admin.ch/dokumentation/energieforschung/](http://www.bfe.admin.ch/dokumentation/energieforschung/)



27. Gary (14 May 2006) Solar wall constructions. <http://www.builditsolar.com/Projects/SpaceHeating/SolarWall/SolarWall.htm>
28. Douglass E (10 Nov 2007) A pioneer refuses to fade away His passion for solar still burns – forty years ago, Harold Hay came up with a way to heat and cool homes using water and the sun. Los Angeles Times, Los Angeles
29. Fachverband Transparente Wärmedämmung e.V (2000) Bestimmung des solaren energiegewinns durch massivwände mit transparenter wärmedämmung. Fachverband Transparente Wärmedämmung e.V, Gundelfingen. [www.umwelt-wand.de](http://www.umwelt-wand.de)
30. Humm O (1998) Transparente wärmedämmung – mehr kollektor als wärmedämmung. Oerlikon Journalisten AG, Zürich
31. Domenig-Meisinger I, Willensdorfer A, Krauss B, Aschauer J, Lang G (2007) Erstes mehrfamilien-passivhaus im altbau passivhausstandard und -komfort in der altbausanierung am beispiel eines großvolumigen MFH in Linz. BMVIT, Vienna. <http://www.nachhaltigwirtschaften.at>
32. Domenig-Meisinger I, Grünewald S, Rottensteiner S (2010) Apartment building on Makartstraße, Linz AT, IEA SHC Task 37 Results. IEA, Linz. [www.iea.shc.org](http://www.iea.shc.org)
33. Hastings R, Morck O (eds) (2000) Solar air systems – a design handbook. Earthscan, London. ISBN 1-873936-86-9
34. Hastings R (ed) (2000) Solar air systems – built examples. Earthscan, London. ISBN 1-873936-85-0
35. Schoch R (1995) Reihenhäuser wydacker, zollikofen. Aarplan Architects/Matter+Amann HVAC, Bern. [www.energienetz.ch/solaregebaeude/Text/SG/ATR/Zollikofen.pdf](http://www.energienetz.ch/solaregebaeude/Text/SG/ATR/Zollikofen.pdf)
36. Hastings R, Wall M (2007) Sustainable solar housing: vol 1, strategies and solutions. ISBN-13: 978-184407-325-2; vol 2, exemplary buildings and technologies. ISBN-13: 978-184407-326-9. Earthscan, London
37. Hanus C, Hastings R (2007) Bauen mit solarenergie. Vdf Hochschulverlag AG, Zürich. ISBN 978-3-7281-3085-3

## Pathogen and Nutrient Transfer Through and Across Agricultural Soils

DAVID M. OLIVER<sup>1,2</sup>, LOUISE A. HEATHWAITE<sup>2</sup>

<sup>1</sup>Biological & Environmental Sciences, School of Natural Sciences, University of Stirling, Stirling, UK

<sup>2</sup>Centre for Sustainable Water Management, Lancaster Environment Centre, Lancaster University, Lancaster, UK

### Article Outline

Glossary

Definition of the Subject

Introduction  
Agriculture, Livestock, Manures, and Contaminants  
The Transfer of Pathogens and Nutrients Through and Across Soils  
Future Directions  
Bibliography

### Glossary

**Hydrological connectivity** The linkage of spatial locations through different hydrological flow paths (surface and subsurface) within the catchment drainage network.

**Farmyard manure** Feces and urine mixed with bedding material (such as straw) used for housed livestock, and recycled back to land as an organic fertilizer.

**Fecal indicator bacteria (FIB)** Nonpathogenic microbial parameters that can be used as surrogate measures of infection risk to humans.

**Leaching** The movement and loss of soluble elements and colloids from soil via drainage water to both surface water and ground water environments.

**Matrix flow** The slow percolation of water through the soil pore system.

**Mobilization** Term used – in the context of this paper – to describe the initiation of contaminant transfer and the process by which those contaminants begin movement from soil.

**Nonpoint source pollution** Comprises contamination and pollution arising from many dispersed sources.

**Pathogens** Microorganisms capable of causing disease or illness in a host and used here to refer to bacteria and protozoa originating from fecal material.

**Preferential flow** Rapid movement of water and contaminants through the soil architecture. Much of the flow is focused in regions of enhanced flux, such as earthworm burrows or larger soil pores (macropore flow).

**Slurry** A liquid mix of feces and urine produced by housed livestock combined with water during management, and usually incorporating some bedding material to give dry matter content of 1–10%.

**Surface runoff** Flow generated from rainfall and other water sources that facilitates the transfer of

contaminants across the soil surface due to saturation excess or infiltration excess conditions.

**Transfer** A term used here to describe the movement of pollutants through soil-water systems.

### Definition of the Subject

Human activity can place heavy stress on agricultural soils across the world. Soil systems are continually manipulated in order to support the increase in crop yields and accommodate more intensive livestock production and thus provide the planet's ever-growing population with a diverse array of ecosystem services, among which food production features highly. The recycling of livestock manures to land provides a sustainable solution to support the ecosystem services that soils provide and a host of benefits both in terms of improving soil structure and also soil fertility. However, livestock manures and feces may contain a high number of fecal microorganisms that pose a threat to human well-being and potentially large concentrations of nutrients harmful to the ecology of freshwater systems that the soils often buffer. To manage water quality in agricultural catchments value should be placed on adopting appropriate land, livestock, and manure management. To do this effectively requires a comprehensive understanding of the role of soil in facilitating contaminant transfer via a suite of hydrological pathways, and its ability to filter or absorb key pathogens or limiting nutrients during their passage through the soil network. Uncertainties in quantifying episodic transfers of contaminants through spatially and temporally dynamic pathways in soils at different scales ensures that multidisciplinary teams of scientists will not be short of challenges posed by soil complexity for decades to come.

### Introduction

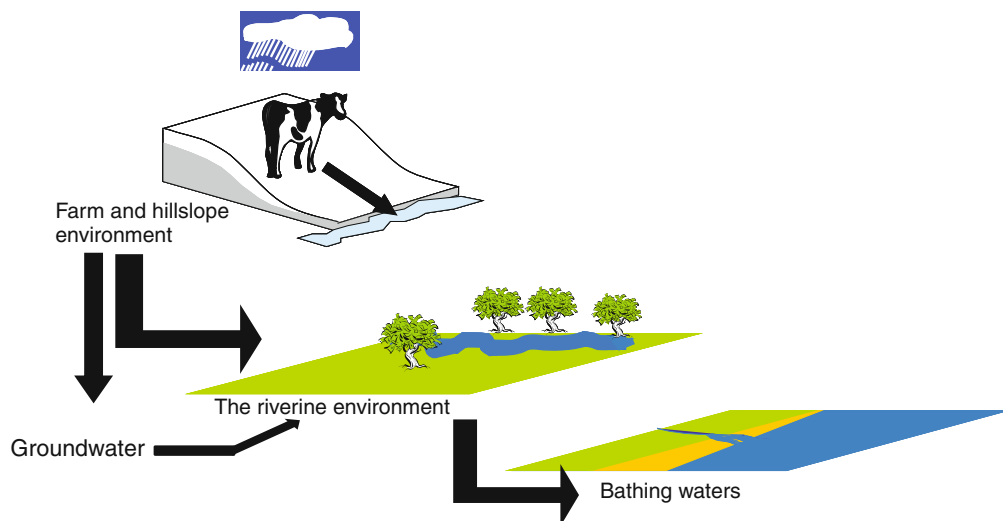
Soils are central to the functioning of agricultural systems and provide the planet's ever-growing population with a diverse array of ecosystem services, among which food production features highly. If managed effectively they can provide a host of benefits to society by, among other things, buffering flooding and hindering the transfer of numerous pollutants [1].

Conversely, if managed inappropriately the consequences can be far reaching, moving beyond the farm boundary to impact on local, regional and national health, livelihoods, food security, and downstream environmental quality. In terms of water quality, this can mean contamination of the ground and surface water supplies with the potential for chronic and acute impacts on human health [2, 3].

In the past nonpoint source pollution from agriculture has been synonymous with nutrient and sediment impacts on water quality. However, there is now a growing recognition that microbial pollution is a significant contributor to water quality impairment across the world [4]. Indeed, pathogen contamination of receiving waters has implications not only for water quality but also human health [5] with pathogens, or indicators of their presence, now topping the list of leading contaminants causing watercourse impairment across the USA [6]. Contamination of surface waters with microbial pollutants and nutrients is therefore potentially problematic both in an ecological and human health context. However, to put this in perspective, while the consequences of large-scale waterborne disease outbreaks linked to agriculture can be fatal, their occurrence is relatively rare. Nonetheless, coastal waters and often lakes and rivers too can be used for bathing, and their deterioration in water quality with fecally derived microbes from agricultural lands provides a key route of recreational exposure of potential pathogens to the human population. The reader is referred to ► [Bioaccumulation/Biomagnifications in Food Chains](#) for a substantial discussion surrounding further impacts of microbial transfer from land to water on food chain contamination. For nutrients, eutrophication is a clear consequence of overloading surface waters with supplies of nitrogen (N) and phosphorus (P). While being aesthetically displeasing, its toxic impacts on freshwater ecology include oxygen depletion leading to fish kills, and cyanobacterial blooms that can prove harmful to humans and animals (e.g., dogs and sheep) [7]. The degree to which eutrophication is considered a problem throughout the world depends on the place and people concerned, but it is an issue for a large number of water quality regulators across many nations.

Pathogen-contaminated groundwater has caused much waterborne disease worldwide [8]. Similarly, nitrate contamination of groundwater supplies is often linked to nonpoint source pollution in agricultural catchments, and it too can cause adverse effects on human health [9]. Waterborne outbreaks of any variety often have a large impact on society, but the actual disease burden attributed from such incidents in Europe is difficult to approximate and most likely underestimated [10]. An example of an outbreak of illness linked to microbial contamination of fresh water supplies is the Walkerton *Escherichia coli* O157:H7 Outbreak, Ontario, Canada in 2000 whereby over 2,500 people became ill through a drinking water well contaminated with farm runoff, and at least seven died directly as a result of drinking the contaminated water. The health implications (both anecdotal and case report data) of exposure to freshwater cyanobacteria are perhaps more contested and debated and have been reviewed recently by Stewart et al. [11]. Protection of surface and groundwater supplies from both nutrients and microbial contaminants thus requires a need to understand the differing transport behavior of these pollutants in agricultural systems, and critically this means understanding the role of soil in providing

a potential buffer between agricultural activity and receiving waters. Thus, there is considerable value in appreciating that the quality of approaches to land management should be reflected in the quality of watercourses that drain catchments. This chapter considers the movement of pathogens through and across soil systems in agricultural settings and draws on nutrient studies to provide a comparative analysis of the differences in the importance of pathway functioning for different contaminant typologies, which ultimately is a key requirement for making measured assessments of the potential for pollution swapping linked to management shifts on-farm. The reader is referred to ► [Microbial Risk Assessment of Pathogens in Water](#) for a discussion of chemicals and pathogens in aquatic systems, and therefore this current chapter does not consider in-stream dynamics of nutrients or pathogens and their response to aquatic conditions when delivered into receiving waters. However, the movement of these biological and chemical contaminants through hillslopes and associated soil matrices is critical in linking on-farm activity to wider downstream impacts at both riverine and coastal environments, but also groundwater supplies too as identified in Fig. 1 (see ► [Microbial Risk Assessment of Pathogens in Water](#)).



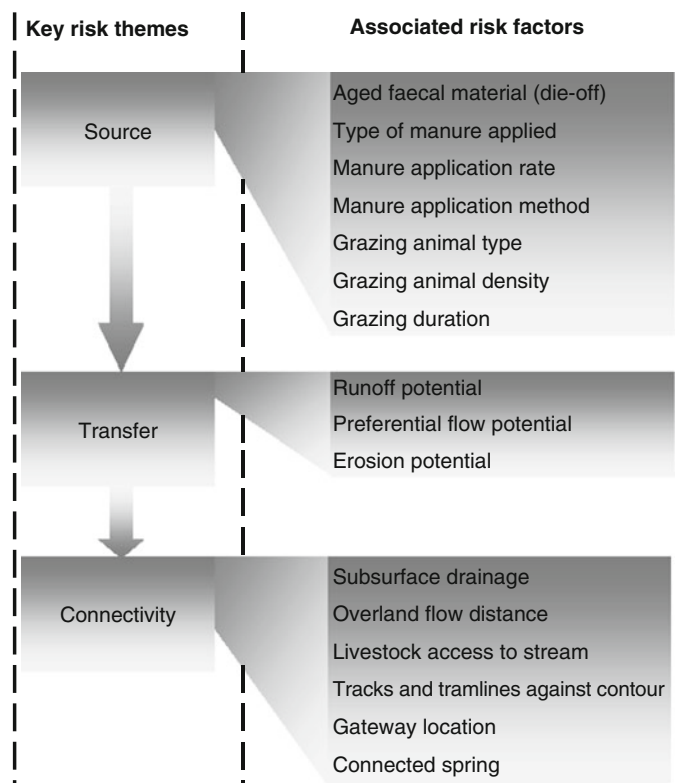
**Pathogen and Nutrient Transfer Through and Across Agricultural Soils. Figure 1**

“Farm gate to the bathing zone”: diverse subsectors of catchments requiring investigation of nutrient and fecally derived microorganism fate and transfer

## Agriculture, Livestock, Manures, and Contaminants

Pasture that receives manure, through either land application in varying forms (e.g., solid, liquid) or via direct deposition by livestock, will accommodate some degree of risk for contributing fecal microorganisms and nutrients to watercourses, as will arable soils (though direct defecation is not an important input for these systems). Many jurisdictions mandate how livestock wastes are managed to protect adjacent water quality from microbial and chemical contaminants that pose an environmental and human health challenge [12]. The likelihood of microbial or nutrient loss from land to water may vary in relative magnitude from negligible through to very high risk potential. This will depend on the combination of a number of site-specific source, transfer and connectivity drivers (see Fig. 2) thought to moderate the risk of contaminant loss from land to water. Landscape features and

management practices help determine this risk potential [13]. In effect, catchment contaminant dynamics depend on a complex interaction between spatial patterns of land use and management, soils, antecedent conditions, and rainfall/event characteristics [14]. For example, different soil types can play a large role in determining the magnitude of transfer of fecal microbes, delivered to land via applied manures, through different soil hydrological pathways. In turn, the varied soil characteristics linked to different soil types can impact on the survivability and viability of microbes once within the soil habitat (either as freely suspended entities, or associated with soil or organic matter). Soils are also critical in facilitating, enhancing, or hindering the transfer of nutrients and fecal microorganisms through the pore architecture. Similar to microbial pollutants, nutrients too are input to land via organic wastes, but are additionally sourced from inorganic fertilizers. Background levels of nutrients



**Pathogen and Nutrient Transfer Through and Across Agricultural Soils. Figure 2**

Risk factors associated with source, transfer, and connectivity of pathogens and nutrients in the farm environment. The in-stream and in-river impacts are considered in ► [Microbial Risk Assessment of Pathogens in Water](#)

within soil reservoirs also provide a secondary source of nutrients that can be mobilized and transferred through the soil system, with a proportion ultimately delivered to receiving waters. Clearly, how land use patterns affect the transport, source, and relative longevity of pathogens (and nutrients) in the environment is a significant issue for policy formulation [15, 16].

### Dangerous Microorganisms in Agricultural Systems

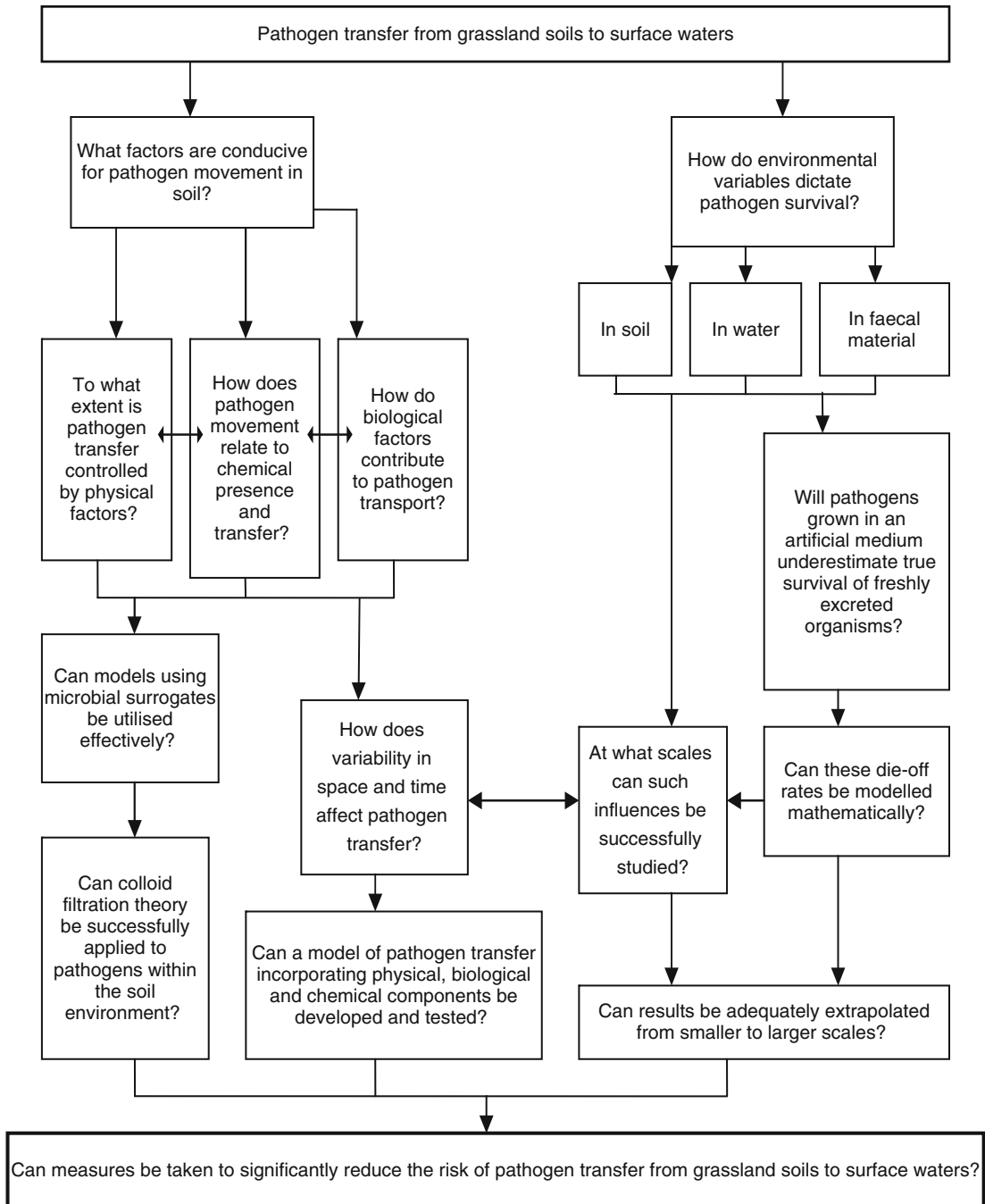
Livestock host a diversity of microbes within their rumen and digestive tract. A proportion of this microflora may comprise pathogenic microorganisms meaning that agricultural systems can harbor reservoirs of bacteria and protozoa potentially harmful to human beings. These reservoirs may be found both within animals but also their excreted fecal material. Thus, microorganisms such as *E. coli* O157, *Salmonella* spp., *Campylobacter jejuni*, *Listeria monocytogenes*, *Cryptosporidium parvum*, and *Giardia intestinalis* may contaminate pasture because of feces excreted during grazing regimes, and due to manures and slurries being applied to land via traditional farming practices. As a result, it is inevitable that the soil system will, at least periodically, harbor fecally derived microbes both at the soil surface and within the network pore structure [17]. Future emerging pathogens will likely arise from existing ones through adaptation and via unpredictable changes linked to biotic and abiotic mechanisms [18].

Fecal indicator bacteria (FIB) are bacteria that are indicative of fecal pollution and are present within all human and livestock feces. They do not pose a significant health hazard themselves; rather they suggest the potential for the presence of pathogenic microorganisms (whether bacterial, protozoan, or viral). The most commonly used FIB are *E. coli*. There are lively debates surrounding the validity of FIB as surrogates for bacterial, protozoan, and viral pathogens [19, 20], but currently FIB are used by many regulatory bodies across the world to monitor microbial water quality. In considering the transfer of microbial pollutants through and across soil there are a number of critical research questions that have driven the continued investigation of this research area. These are summarized in Fig. 3. Clearly there is a need to account not only for microbial transfer, but simultaneously appreciate their survivability when excreted into an

environment outside of the animal gut. If survivability is so unlikely under the environmental conditions typical of agricultural settings then the likelihood for mobilization and delivery to watercourses and downstream impact is much reduced. That said, many microbial pathogens and associated indicators are able to survive for lengthy periods outside of the gut, meaning that there is a key need to understand how they transfer in addition to how their population profiles may change through time when associated with different environmental matrices. The entry ► [Recreational Water Risk: Pathogens and Fecal Indicators](#) provides a comprehensive overview of monitoring approaches used in the detection of pathogens and FIB in the environment.

### The Nutrients of Concern

The primary nutrients of concern considered in this chapter with regard to movement through soils are phosphorus (P) and nitrogen (N), specifically nitrate ( $\text{NO}_3^-$ ). Both P and N are key limiting nutrients in aquatic systems. Contamination of surface and groundwater with anthropogenic inputs of N and P following their passage through and across soils risks the quality of drinking water supplies and may lead to excessive harmful algal blooms [21]. Nutrients are introduced as a comparative contaminant within this Chapter to highlight differences in terms of their transfer dynamics in relation to microbial transfers. The FIB and pathogens discussed within this chapter generally range in size from 1 to 5  $\mu\text{m}$  and can be classified as biological “particles.” In contrast, phosphorus can be divided into a particulate and soluble fraction (soluble in this instance being based on an artificial laboratory subdivision of 0.45  $\mu\text{m}$ ), and  $\text{NO}_3^-$  is highly soluble. This intuitively leads to clear differences in modes of transfer linked to physical mobilization by water for particulate contaminant typologies, in contrast to dilution effects associated with increased discharge volume for soluble contaminants and therefore suggests that there are differences in transfer linked to contaminant functional type. Another important property of P that can impact on its transfer dynamics is related to its high capacity for sorption to soil relative to  $\text{NO}_3^-$  and its association as colloidal P. Indeed, in soils  $\text{NO}_3^-$  remains free in soil solution because the nitrate anion is not adsorbed onto soil surfaces due to soil



**Pathogen and Nutrient Transfer Through and Across Agricultural Soils. Figure 3**

A questioning framework of key research needs for pathogen transfer studies in the environment (Reproduced from Oliver et al. [17])

particles having a partial negative surface charge. As a result,  $\text{NO}_3^-$  is highly mobile in soils. In contrast, ammonium ( $\text{NH}_4^+$ ) is positively charged and can therefore attach and associate with soil particles to be retained within the soil body. However, this chapter only provides an introduction to the differing properties of  $\text{NO}_3^-$  and P relative to microbial contaminants and much more comprehensive detail specifically on nutrient mobility in agricultural systems can be found within recent comprehensive reviews of Heathwaite [22], Edmeades [23], and Haygarth et al. [24].

### Sources of Pathogens, Indicators, and Nutrients

The presence of pathogens, FIB and nutrients within the farm environment can be largely attributed to either point or nonpoint sources. Point sources are readily identifiable inputs that can be traced to a point of origin. They represent “end-of-pipe” locations and can include agricultural ditches, leaking septic tanks and manure stores, and farm hard standings. Nonpoint sources relate to inputs that occur over a large area and are attributed to land use. Nonpoint sources can include dung deposited to pasture by grazing animals [25, 26] and liquid and solid manures spread to land [27]. In addition, treated human wastes can be recycled back to land (biosolids) and still may contain a number of fecal microorganisms of concern. Untreated human waste (i.e., waste from septic tanks) may sometimes be spread back to land, though in the UK this is against regulatory policy. However, anecdotal evidence suggests this activity does take place, though published data on such practice are notably scarce. Wildlife can also serve as a source of microbial pollutants and indeed additional P and N inputs to land via their excretions. However, the significance and quantification of the wildlife source component is an often overlooked contributor to pasture-based budgets [28].

The soil matrix can sustain fecal microorganism survival, especially if microbes are incorporated in association with fecal material. Indeed some studies have detailed bacterial cell survival in excess of several months [29]. This can be due to protective microsites and protection from desiccation and UV radiation. Clearly then, it is not only fresh additions of manures to land that may be capable of impacting the microbial quality of watercourses, and aged fecal remnants can still contribute bacterial numbers to runoff following

rainfall [30]. Slurries, solid manures, and feces all harbor different numbers of fecal microbes, and all impact on differential fecal microbe die-off patterns [31]. Furthermore, the consistency of each manure type (e.g., dry matter content) allows for differing degrees of mobility of fecal microbes accommodated within these manure source locations [32].

Nutrient sources in agricultural systems are somewhat different in that they do not depend on a fecal store to convey signal strength in receiving waters. For example, natural sources of N and P exist in the environment, and for N there are atmospheric sources too. The soil can therefore serve as a reservoir for N and P, meaning that even without continued anthropogenic inputs to land there will still be considerable losses of  $\text{NO}_3^-$  and P from land to water following rainfall events. This is a critical difference between the fecally derived bacterial and protozoan contaminant typologies and nutrient contaminant typologies. While microbial contaminants can persist for lengthy periods, there are not significant soil stores of fecal pathogens and their indicators *independent* of manure applications or livestock activity, though some studies report the potential for naturalized *E. coli* populations in the soil environment [33–35]. In contrast, liver fluke and intestinal worms and similar livestock pathogens do have significant soil stores that are important for their cycling, but they are not considered in this chapter under the definition of pathogens (e.g., fecal bacteria or fecal protozoa). Therefore, lag time between cessation of contaminant inputs to land and a measured reduction of pollutant losses at the edge-of-field scale varies by the pollutant type and depends strongly on the behavior of the pollution source [36]. Circumstances where farming practices have led to excessive soil P levels can be particularly problematic because even if nutrient management leads to a reduction of P inputs to levels below crop removal rates, the timescale needed to exhaust the P from the soil to the point where dissolved P in runoff is effectively reduced may well be years to decades [36]. Of course, soil with a high P index (high P content) does not always result in high P transport as this is dependent on the way that the P is “locked” into the soil and its likelihood for subsequent mobilization.

Manure and slurries are applied to land via a range of techniques and equipment. Broadcast-applied manures (i.e., those spread to land using a splash

plate and distributed onto the surface of blades of grass) are likely to be considered more risky in terms of contaminant loss than injected or plowed manures because of the opportunity for incidental contaminant losses from land to water [37]. However, broadcast applied slurry is likely to lead to a more rapid destruction of associated bacteria through UV radiation and desiccation [38]. Livestock grazing is equally, if not more important than manure applications in loading the landscape with fecal microbes [39]. The key difference between land applied manures and livestock excretions is the resulting spatial patterning of contaminants across the landscape. Similarly for nutrients, in this case  $\text{NO}_3^-$ , McGechan and Topp [40] have found higher levels of  $\text{NO}_3^-$  pollution in tile drains following grazing compared to fields receiving slurry and have suggested that high levels of  $\text{NO}_3^-$  pollution could be attributed to various factors, including the fact that cows tend to congregate in certain areas of a field at a localized stocking rate much higher than the overall stocking rate, and due to deposition of N at times when grass cannot utilize it as a plant nutrient. Animal type is a critical factor to consider with regard to the source strength of contaminants in farmed environments. This is because age and variety of livestock governs the number of organisms per gram of feces excreted onto pasture (e.g., lambs shed higher numbers of *E. coli* than beef cattle) but also dictates the daily excretion rates of feces (and therefore nutrients and microbes contained within) to farmed land and impacts on the differential release of FIB from various types of livestock feces at different times of the year. Logically, grazing density bears an impact on land-based loading of fecal microorganisms and nutrients too, and it has been shown that the concentration of FIB in streams in sub-catchments with high stocking densities can be four to eight times higher compared to sub-catchments accommodating low stocking densities [41], and grazing duration is potentially important for both nutrient and microbial contaminants based on the accumulating reservoirs of fecal excretion to land through time.

### The Transfer of Pathogens and Nutrients Through and Across Soils

The transfer of water across and through the soil environment, and contaminants dissolved or entrained

within, is subject to a series of spatial and temporal controls that dictate this potential for transfer and which vary both spatially and temporally [42]. Temporal controls on microbial and nutrient transfers include meteorological inputs that provide an energy source and driving force to initiate transfer processes. Additionally, hydrological pathways are susceptible to change through time, for example, pore networks can collapse, drain pathways can silt up, and extreme events can remove existing pathways and reconfigure hydrological conduits. The significance of a particular pathway in facilitating microbial and chemical transfers from land to receiving waters is controlled by properties of the pollutant of concern and also the form of the contaminant, for example, whether it be particulate, particle associated, or soluble (i.e., the contaminant functional type) [43]. Climate and weather too are key dictating factors controlling the magnitude of transferred loads. Whether considering individual storm events or interannual variability, these climatic drivers can have a significant effect on the magnitude of multiple pollutants over many timescales. Spatial factors impacting on transfer include the particular soil type that microbes or nutrients are negotiating their passage through or across. Soil types dictate hydrological pathways, some being more likely to hinder transfer via filtration, whereas others may be more likely to permit rapid transfers via large cracks or pores in the soil (macropore flow). Research has demonstrated the importance of water flow velocity and soil particle size distributions, for example, on *E. coli* transport through soil [44].

For transfers across the soil surface, it would appear logical to assume that slope gradient and slope shape can be useful indicators for transfer potential, and many models accommodate a slope term as a control over particle transfer, with increased transfer associated with increased slopes. However, some studies suggest the need for caution when dealing with what appears to be a relatively straightforward assumption. Such relationships tend to be based on data collected over a wide range of slopes and using relatively small soil flumes. Recent work by Armstrong and Quinton [45] used laboratory rainfall simulation on a large soil flume to investigate interrill soil erosion of a silt loam under a rainfall intensity of  $47 \text{ mm h}^{-1}$  on 3%, 6%, and 9% slopes. Results from replicate experiments showed wide

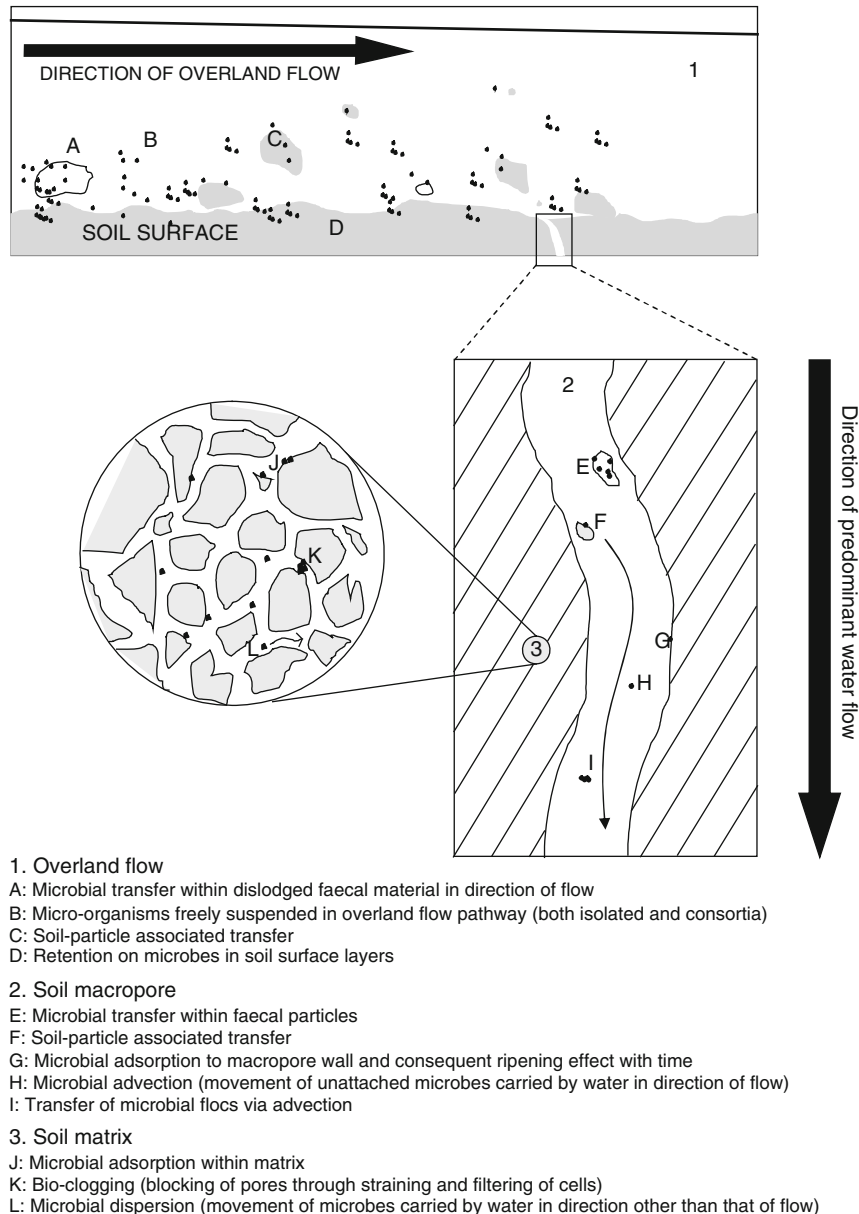


variations in runoff and sediment concentration that were explained by the complexities in interrill soil erosion processes. Critically, the data also demonstrated that at low slope (on arable soils) processes related to surface area connectivity, soil saturation, flow patterns, and water depth may dominate over those related to gravity, and Armstrong and Quinton [45] therefore query the validity of risk assessments and soil erosion models with a dominant slope term when assessing soil erosion from agricultural land at low slopes. However, it is important to acknowledge the scale at which this reported study was conducted, and caveats should be applied in upscaling such laboratory-based findings to field situations. For more topographically diverse landscapes, alternative approaches use landscape wetness as a metric of delivery potential by integrating the small-scale spatial variation in runoff generation and probability of hydrological connection linked to topography [46, 47].

Physical, geochemical, and biological processes can be used to classify the key transport mechanisms responsible for microorganism and nutrient movement within soils. The physical processes include advection, whereby pathogens or FIB and nutrients are carried in bulk water and move according to the water velocity, and dispersion, which can involve the spreading of microorganisms and nutrients as they move along the water path. Geochemical processes act to delay microbial and particulate transfer through the soil matrix and consist of filtration, sorption, and sedimentation mechanisms [48, 49]. Finally, biological processes, such as growth and chemotactic responses, may influence microbial transfer through the soil habitat, albeit to a lesser extent [50, 51]. This ability to self-propel through the soil system because of their biological attributes (e.g., pili, fimbriae) differentiates microbes from their nutrient counterparts. While at the pore scale this transfer mechanism is important, in the grander scheme of the catchment or hillslope, the role of biological transfer is likely to be minimal in contrast to the physical mobilization facilitated by rainfall and the resulting water flows.

The transfer of agriculturally sourced contaminants from soils to ground and surface waters is largely driven by rainfall and resulting surface and subsurface runoff. Conceptually, transfers can be classified as low energy or high energy. For example, slow flow microbial and

nutrient transfers may operate between storm events and are thought to be associated with the steady percolation of precipitation inputs through the soil profile. This contrasts with overland flows and bypass flows resulting from high-energy precipitation (or storm) events, which enable the physical movement of soils, manures, and potential pathogens into streams, creating a more rapid and direct transfer route. However, it is interesting to draw attention to findings from Sharpley et al.'s [52] recent study that showed that small storms were more important than large storms in delivering P to watercourses over a 10-year research period. These two contrasting energy levels of transfer will be discussed in later sections of this chapter and are illustrated in Fig. 4. Nitrate (and some forms of P), being soluble, follow a different trend than that of microorganisms and particulate P, and instead can be diluted under heavy water flow [53]. Table 1 is provided as a point of reference to direct the reader to scale-appropriate studies of the transfer of a variety of agricultural contaminants through soil systems. Current understanding of the transfer capability of different hydrological pathways in space and time is growing for microbial pollutants, but this remains a complex and underexploited interdisciplinary research topic necessitating the collaboration of soil hydrologists and environmental microbiologists. While understanding the spatial and temporal intricacies of environmental transfer of fecal microorganisms lags some way behind current knowledge surrounding N and P dynamics in agricultural systems, that does not mean the knowledge of P and N dynamics is satisfactory. There remains for these nutrients need to further the understanding of the relative importance of different hydrological pathways in providing *delivery* (in quantifiable terms) from land to water [95]. Some have argued for the critical development of appropriate methods to quantify the delivery process and thus new monitoring tools capable of providing a framework for understanding contaminant transfer and delivery at a range of scales in agricultural catchments [88]. Deasy et al. [88] highlighted the dominant role of a field drain in transferring P from land to water and argued that overland flow inputs, despite being directly connected to the stream and containing higher P concentrations, contributed less to the stream P flux. Such data are key reminders that the obvious and visible routes of



#### Pathogen and Nutrient Transfer Through and Across Agricultural Soils. Figure 4

Natural transfer pathways available to fecally derived microorganisms applied to soil surfaces (Reproduced from Oliver et al. [17])

transfer do not always equate into dominant and critical flow pathways upon which mitigation efforts must be focused to protect our water resources. The P transfer continuum concept has been outlined by Haygarth et al. [96] as a framework within which the factors that contribute to nonpoint P fluxes from

catchments can be operationally divided into *source*, *mobilization*, and *delivery*. These divisions are a convenient way of grouping processes that determine P behavior and accommodate transferability in terms of conceptualizing other contaminants in agricultural systems (cf. Fig. 2).

**Pathogen and Nutrient Transfer Through and Across Agricultural Soils. Table 1** A “look-up” table of research studies investigating microbial and nutrient transfer through and across soils at a range of different scales

Scale of study	Contaminant type			
	Bacterial	Protozoan	P	N
Laboratory soil columns/boxes	Artz et al. [54] Donnison and Ross [58] Garbrecht et al. [44] Horswell et al. [64] Rosa et al. [65]	Mawdsley et al. [55] Boyer et al. [59] Harter et al. [62]	Matula [56] Tarkalson and Leytem [60] Brock et al. [63]	Entry et al. [57] Miller et al. [61]
Vertical Lysimeters	Bech et al. [66] Brennan et al. [34] Guzman et al. [69]		Turner and Haygarth [67]	Di et al. [68]
Soil plot : small	Muirhead et al. [70] Abu Ashour and Lee [72] Soupir et al. [75] Mishra et al. [77] Thiagarajan et al. [79] Oliver et al. [82] Kouznetsov et al. [84] Collins et al. [86]	Ferguson et al. [71] Ramirez et al. [80]	Preedy et al. [37] Tunney et al. [73] Quinton et al. [74] Withers et al. [76] Heathwaite et al. [78] Haygarth et al. [81]	Alfaro et al. [83] Olson et al. [85]
Hillslope Sub-catchment - Catchment	Kay et al., [87] Close et al. [90] Davies-Colley et al. [93] McKergow and Davies-Colley [14]	Keeley and Faulconer, [15]	Deasy et al. [42, 88] Heathwaite and Johnes [91] Rothwell et al. [94]	Botter et al. [89] Dougherty et al. [92] Heathwaite and Johnes [91]

### Soil Characteristics Impacting on Transfer

Rainfall and resulting runoff and drainage from farmed land is critical for the transfer of agricultural pollutants such as phosphorus,  $\text{NO}_3^-$  and sediments from land to water (e.g., [37, 97]), and there are a series of key soil characteristics that impact on the transfer of pathogens and FIB through lysimeters, hillslopes, and catchments too. Soil water content, soil structure, and soil texture are among those that shall be discussed within this chapter. The physiochemical properties of soils and the resulting interactions with potential pollutants means that the navigation of pathogens and nutrients through the soil architecture is extremely complex, and particular attention will be given to these rapid versus slow transfers as dictated by soil properties in the following sections of the chapter. Particulate-type contaminants experience more rapid travel times in coarser textured soils with larger pore spaces as opposed to finer textured soils, with the soil matrix operating as a filtration system. In soils where matrix flow

dominates it is therefore generally accepted that water-induced particle transport is strongly correlated with particle size.

In addition to the physical make-up of soil and the associated network of conduits, pores, and microhabitats, it is pertinent to appreciate the variability in sorption properties attributed to different soil types. This property of soil and associated colloidal material can have a marked influence on microbial and P transfer [98, 99]. The major soil components affecting sorption of bacteria and P are clay and organic matter, and particle and colloid facilitated TP delivery from soils to water via different hydrological pathways has been shown to be dominated by the transfer of TP associated with clay and colloidal fractions [81]. For *Cryptosporidium*, hydrophobicity and zeta potential have been highlighted to exert a significant influence in the adhesion mechanisms of the oocysts [100], and more recently the presence of manure in solution has been shown to enhance the extent of adhesion of these

protozoa to soil particles. However, the role of manure in facilitating attachment of *Cryptosporidium* to soil particles is complex, with indications that an optimal concentration of a “facilitating” component of the manure exists (at somewhere between 0% and 1.0%) [101]. Clearly, the sorption of microorganisms and nutrients to soil surfaces cannot be attributed to a single factor; instead it is important to appreciate an array of forces interacting to govern microbial retention. Table 2 provides a definition of the common mechanisms through which FIB and bacterial pathogens associate with soil particles.

From a conventional soil science perspective the concept of sorption is normally applied to chemical contaminants and the derivation of their associated adsorption isotherm. Elements of sorption theory can be considered transferable for the determination of bacterial association with soil. Adsorption is an important process that partly governs the aqueous concentration of contaminants in soil environments, and therefore their mobility. The common approach is to use a batch equilibration method to obtain adsorption coefficients. This is based on the assumption that partitioning of the contaminant is a thermodynamically motivated distribution between two readily separable, homogeneous phases [102]. Many other studies have investigated the adsorption of a broad range of

potential contaminants, including heavy metals, pesticides, herbicides, and fertilizers.

The adsorption isotherms of nutrients such as P often fit models such as those described by the single and double Langmuir, Freundlich, and Tempkin equations whose profiles suggest that a maximum loading of P with the soil occurs, after which no more sorbate may be adsorbed. The linear isotherm is essentially a special case of the Freundlich isotherm, and this suggests that no point is reached whereby the solid fraction has had all its available sorption sites exhausted. Oliver [103] observed a linear isotherm when investigating the affinity of *E. coli* to soil particles suggesting that sites for cell interactions remain available. However, the linearity of the data may in fact reflect multilayer cellular associations with particles whereby cell consortia have associated with a contact point of the particle. Koopmans et al. [104] discuss isotherm linearity with respect to P adsorption and propose that a linear profile is often either: (1) a function of a narrow range of concentrations being tested; or (2) linked to the use of a wide soil to solution ratio. In the *E. coli* isotherm described by Oliver [103] neither of the two factors discussed by Koopmans et al. [104] were thought to apply to the data because the range of concentrations used was large in terms of a realistic minima and maxima, and an appropriate soil to water ratio was determined in one

**Pathogen and Nutrient Transfer Through and Across Agricultural Soils. Table 2** Summary table of mechanisms of bacterial sorption/attachment to soil particles

Potential mechanism of sorption/attachment	Description
Van der Waal forces of attraction	Cells overcome electrostatic repulsion and are held at a finite distance from a particle surface
Charge attraction of opposite signs	Metal oxide coatings on soil particles confer a positive charge to the particle surface and this results in a much tighter adhesion of the cell to the surface
Hydrophobic forces	Hydrophobic bacteria have the tendency to adsorb to surfaces in water due to repulsion from the polar water molecule
Positive chemotaxis (cellular mobility)	Higher nutrient content of a particle surface may cause motile bacteria to move toward surface as a result of nutrient gradient
Irreversible anchoring via pili and fimbriae	Biological mechanism of attachment where contact is made between cell and surface via hair-like appendages. The presence of a localized positive charge at the tip of fimbriae can aid attachment to a negatively charged soil particle
Charge fluctuations	Heterogeneities in surface charge provide locally favorable regions for attachment

of the preliminary stages of designing the isotherm experiment. Of course, if the concentration of cells was increased further then the *E. coli* isotherm may reach a plateau at a potential retention maximum, but the applicability of such results to field situations is likely to be limited. While some studies have determined that over 99% of cells adsorb to soil [105], the numbers do not necessarily equate to real-world observations whereby a greater proportion than the remaining 1% of total applied cells loaded onto the soil surface via manure and feces may emerge via runoff and subsurface drainage. The reasons for this are numerous. For starters, hydrological factors play a role at larger scales and are unaccounted for in laboratory batch-scale experiments. Essentially, water flux may prevent cell interaction with the soil because preferential and rapid overland flow pathways can allow bacteria to bypass the soil matrix and those soil horizons that adsorb contaminants strongest. Additionally, bacteria entering the soil system under field conditions are not necessarily free-living cells as introduced to the soil-water system under experimental conditions. Fecal bacteria may already be associated with organic matter derived from excrement, and this may impact on soil-cell interactions. The extent to which laboratory findings can be extrapolated to the environment is, therefore, a little unclear. However, such comparative studies can provide a mechanistic understanding of an important process and as long as their limitations are acknowledged such experiments are important tools in the development of the understanding of soil-contaminant interactions.

### Colloid and Particle Associated Transfer

In addition to retention of microbial and nutrient contaminants within the soil architecture there also exist the potential for these contaminants to associate with colloidal and particulate soil material, which may subsequently provide a vehicle for contaminant transfer through the pore network. There are two particularly important properties associated with colloids that enable them to function as important contaminant carriers. The first is that colloids have a very large specific surface area, in excess of  $10 \text{ m}^2 \text{ g}^{-1}$  [106]. Second, these colloids remain stable in suspension for significant periods, and if bacteria attach or nutrients such as P adsorb to the large available surface area, their

dispersal through the soil may be aided considerably. Those colloids that remain more in the center stream of the flow path are likely to remain uncaptured and thus migrate along these faster, more permeable flow paths.

The strong positive relationships between suspended sediments and P in drainage waters reported throughout the literature demonstrate the clear linkage between this nutrient and soil particles. Much work has considered colloidal P transfer through agricultural systems, and it is an accepted mechanism by which a significant fraction of P loss from agricultural land can occur. Colloidal material falling within the notional “dissolved” fraction (i.e.,  $<0.45 \mu\text{m}$ ) has thus been shown to serve as an important vehicle for the transfer of P from soils [78]. In fact, a whole range of particle and colloid size fractions have been shown to function as P carriers with linear regression highlighting strong significant relationships ( $r^2 = 0.86$ ) between water extractable supernatant turbidity and colloidal P release across particle sizes ranging between 2 and  $0.0003 \mu\text{m}$  for a suite of different soil types [78]. The association of P with these colloidal fractions allows for a relatively rapid transfer of sorbed material because of the exclusion of colloidal size fractions from soil micropores and restricts their passage to the more rapid flow routes via macropores.

Association of FIB with soil particles has been discussed too (e.g., [72, 107–109]), though there remains a large degree of uncertainty regarding “vehicles” for cell transfer under environmental settings and a clear need for continued research to consolidate understanding of processes governing cell partitioning with different sedimentary fractions and the importance of such interactions in freshwater systems [5]. Importantly, the size of *E. coli* itself is generally regarded as being approximately  $2 \mu\text{m}$  in length and  $0.5 \mu\text{m}$  wide. In one of the few studies examining the preferential attachment of *E. coli* to different particle size fractions of an agricultural grassland soil [108] it was shown that 35% of introduced *E. coli* cells were associated with soil particulates  $>2 \mu\text{m}$  diameter. Of this 35%, most of the *E. coli* (14%) were found to be associated with the size fraction  $15\text{--}4 \mu\text{m}$ . This was attributed to the larger number of particles within this size range and its consequently greater surface area available for attachment. When results were *normalized* with respect to estimates of the surface area

available for bacterial cell attachment to each size fraction, it was found that *E. coli* preferentially attached to those soil particles within the size range 30–16  $\mu\text{m}$ . For soil particles  $>2 \mu\text{m}$ , *E. coli* showed at least 3.9 times more preference to associate with the 30–16  $\mu\text{m}$  than any other fraction apart, from the  $<2 \mu\text{m}$  grouping. Undoubtedly the range of methods used to arrive at estimates of bacterial attachment to suspended sediments can in many respects be responsible for the discrepancies observed in different studies [110]. Within emerging drainage water, a measure of turbidity should be correlated with FIB concentrations if the majority of cells are truly associated with suspended solids and, in addition, such association is unaffected by other water quality parameters. However, both strong and very weak relationships have been observed by various authors between *E. coli* concentrations and turbidity, and between *E. coli* concentrations and total suspended solids concentrations [110]. This is perhaps not too surprising given that the microbial consortia being measured reflect a combination of life-cycle stages. For example, turbid water can often be found draining pasture that has not been grazed for several months. The length of time between the removal of livestock and the collection of runoff water can mean that the majority of FIB may have perished as a result of UV radiation and desiccation and so the relationship between turbidity and FIB concentrations is weak. In contrast, a storm event that generates highly turbid runoff from pasture grazed by a large number of cattle will likely result in a strong relationship between the suspended particle concentration and FIB numbers because the population of FIB present are abundant within freshly deposited material. Tracking the actual emergence of FIB, pathogens, and turbidity peaks with storm hydrographs provides one interesting means of analyzing the strength of association of microbial contaminants with the soil fraction. Such observation as reported by Oliver et al. [103] identified a similar but delayed emergence pattern of *E. coli* and turbidity peaks in subsurface runoff via artificial drainage from grassland plots. Thus, the peaks for turbidity, flow, and *E. coli* concentration were not identical, but occurred in respective order as a function of increasing time. The results suggested that *E. coli* may have been associated with hydrologically energized soil particles of a particular size fraction.

### Rapid Pathways of Contaminant Transfer

Intensity of water flow through soils is undisputedly one of the most critical factors responsible for driving the transfer of fecal microbes and nutrients from land to receiving waterbodies. However, when entering receiving waters that may have a high discharge there is the potential for high dilution capacity. Surface runoff processes that generate overland flow are often perceived to be the dominant transfer pathway for pathogens and microbial contaminants. The growing evidence base detailing pathogen and FIB transfers through agricultural systems is beginning to add weight to such hypotheses. If antecedent soil conditions are conducive to generate overland flow and heavy rains occur shortly after slurry application or an intensive grazing season, then there is potential for significant runoff of fecal microorganisms and particulate P following their entrainment into a variety of flow pathways. Overland flow that directly connects with a watercourse has been shown to deliver substantial loads of FIB to the stream network [86], though the numbers may be diluted if entering a receiving water with high discharge. It is likely that a greater *uninterrupted* overland flow distance and contributing area will enhance the potential for accumulated carriage and delivery of FIB to the watercourse. However, there is little proof that overland flow, once started, actually delivers contaminants to streams and primary water systems in a single event. It may instead provide a pulsing mechanism of transfer or be interrupted by buffers before having impact on receiving waters. Alternatively, particulate contaminants (e.g., cells or particulate P) that are transferred via the surface runoff pathway may be deposited and resuspended numerous times before being finally delivered to a watercourse. The role of high rainfall and storm events and the resulting flow signatures in dictating microbial transfer from soil to water has been documented by a growing number of studies. Recent examples include: McKergow and Davies-Colley [14], Sinclair et al. [111], Wilkes et al. [20], Davies-Colley et al. [93], and Oliver et al. [82]. Similarly for P and N a large amount of literature is available on the response of these nutrients to hydrological drivers (e.g., [53, 78, 89, 112]). Some recent studies on P delivery from land to water have however challenged prior assumptions of the

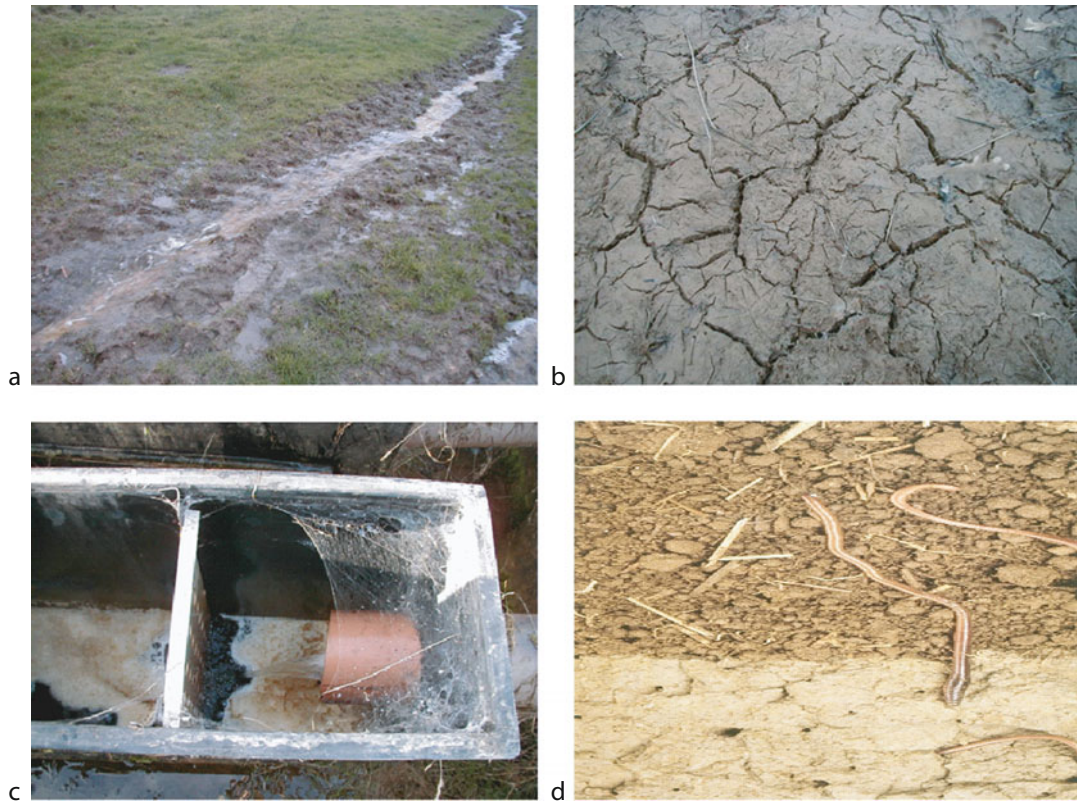
importance of large storm events [52, 113]. Sharpley et al. [88] compared the surface runoff contributing area and stream flow and total P response for 248 storms over a 10-year period from 1997 through to 2006 and found that 93% of storm flows had a return period of <1 year and delivered 63% of the flow and 47% of the total P load. Similarly Jordan et al. [113] were able to demonstrate the importance of high resolution sampling frequency on detecting P delivery to water from diffuse sources independent of storm events and others have recognized the need to adopt appropriate time intervals and lengthy data records to observe useful diffuse pollution trends across scales [22].

Coarser textured soils tend to allow for rapid vertical transfer of  $\text{NO}_3^-$  relative to fine textured soils, though it must be acknowledged that rapid  $\text{NO}_3^-$  transfers can occur within finer textured soils via cracks or fissures, particularly if  $\text{NO}_3^-$  is heavily loaded onto the soil surface. Overall,  $\text{NO}_3^-$  losses via rapid overland flow pathways are relatively minor compared with  $\text{NO}_3^-$  lost via leaching and drainage. For a soluble contaminant such as  $\text{NO}_3^-$  it is interesting to observe response trends over both seasonal and storm-event timescales. For example,  $\text{NO}_3^-$  emergence generally reduces during the growing season and increases during the winter owing to uptake from plants during the summer. However, superimposed on this seasonal pattern of increasing and decreasing concentrations are dilution effects attributed to storm events whereby heavy rainfall effectively causes a reduction in concentration through increased water volume. Downscaling from seasonal timeframes to individual storm-events, the relationship between flow and  $\text{NO}_3^-$  concentration is in fact much more complex and not a straightforward dilution for every rain event.

For microbial transfer, Tyrrel and Quinton [114] summarize overland flow transport scenarios as (1) incorporation of free microbes into overland flow; (2) mobilization of soil or waste particles into overland flow carrying attached microbes; and (3) detachment of microbes from soil surfaces arising from shearing forces of raindrop or flow action. However, there remains much work to be done in quantifying the efficiency of overland flow in facilitating the wash-in of fecal material from pasture to stream. The complexity of predicting contaminant delivery to water courses

cannot be understated, and while wash-in of fecal matter has long been recognized as a consequence of overland flows generated within the contributing areas of a catchment [115], within large and complex watersheds the bacteriological and chemical quality of a stream is the resultant effect of a variety of indistinguishable sources, and so determination of the loading capability of a particular transfer route is difficult. Poaching and pugging, which are terms used to refer to the compaction and breakup of soil due to trampling by livestock, can lead to a lower soil infiltration rate [116] and can enhance the initiation of overland flow. Poaching and pugging are therefore of critical importance in localized areas and hence may be more important at such scales within surface water dominated catchments compared to groundwater dominated catchments, where their effects are likely to be more hydrologically isolated [117]. Despite the apparent complexity in understanding contaminant transfers through soil systems at catchment scales some studies suggest that complex behavior patterns can be reduced to surprisingly low variability in model outputs [118].

Rapid routes of water transfer other than overland flow do exist and include macropores and artificial drainage systems, and a suite of rapid flow pathways are illustrated in Fig. 5. Such bypass routes within the soil matrix facilitate a rapid transfer of water and the soluble and particulate contaminants carried within. The significance of large pores and voids in facilitating the movement of water and colloidal material through the soil has long been acknowledged [119–121]. Macropores may be formed naturally or through soil fauna activity, plant root presence, or soil shrinkage. Preferential flow in soil has both environmental and human health implications since it favors contaminant transport to groundwater without interaction with the chemically and biologically reactive upper layer of soil [122]. The interconnected pore network therefore allows for a slow transfer of particulate-type contaminants via infiltration into the soil matrix coupled with a rapid transfer of microbes and particles (and solutes) within larger sized pores capable of increased soil water velocities. In the absence of macropores the distribution of microbes may be mostly confined to the uppermost zones of the soil profile [54]. However, for FIB, after a dry spell, the rapid entry of cells into soil following rainfall does not guarantee rapid downward



**Pathogen and Nutrient Transfer Through and Across Agricultural Soils. Figure 5**

Rapid transfer pathways across and through soils: **(a)** overland flow conduits across a clay loam soil; **(b)** soil cracking on exposed grassland soil surface for a clay loam; **(c)** drain flow exported from mole and tile drains on clay loam grassland plots; **(d)** creation of macropores within soil by earthworms

transport because of the potential for cellular retention in the upper dry topsoil aggregates [123]. Rather, as soils wet up over time the interaction with wet soil particles reduces and cells are able to transfer more freely. Thus, Guber et al. [123] showed that FIB were able to associate with dry soil aggregates in increased number compared to wet aggregates, leading the authors to propose an interesting interplay between soil water content prior to a rainfall event and the subsequent transfer of manure-borne bacteria with infiltrating rainfall. Similarly it has been shown that fecal bacteria sticking efficiency to quartz sand particles decreases with distance traveled [124].

Although macropores often make up only a small volume of the soil body they can serve as vertical and lateral routes of relatively rapid water flow and allow microorganisms, among other colloids and

contaminants, to successfully bypass the sieving and constraining architecture of the soil matrix (see Fig. 4). Attempts have been made to develop relationships between soil type classes in New Zealand and breakthrough curves [125] and to therefore generate datasets of regionalized potential for microbial bypass flow based on soil classifications that become useful for larger scale modeling scenarios. The role of continuous macropores within silt loam and loam soils has been well documented by Abu-Ashour et al. [120]. Their experimental protocol was designed to create artificial macropores (vertical and straight) through sieved and packed soil. A comparison was made between columns (175 mm high x 89 mm diameter) accommodating artificial macropores (and columns without) in their ability to transfer a slug of cell suspension to depth. Irrigation was applied 24 h after inoculation at a rate of



60 mL h<sup>-1</sup> for 2 h. For all packed soils without macropores no biotracer was ever detected in the effluent, regardless of initial water content, rainfall rate, or soil type. The study concluded that initial soil moisture appeared to have a notable effect on bacterial movement through soils especially in the presence of a macropore. The effect of the macropore was not substantial when the soil was dry. Aislabie et al. [126] compared four contrasting soil types for macroporous transfer of bacteria and showed the importance of macropore flow in breakthrough curves of clayey soils. This complements the studies of Paterson et al. [127] and Mawdsley et al. [65], who also demonstrated greater recoveries of introduced microbes in leachates from clay loam compared with loamy sand cores. The study of McLeod et al. [128] observed FIB numbers emerging from saturated intact columns (500 × 500 mm) of 2 contrasting soil types following dairy shed effluent application and simulated rainfall (5 mm h<sup>-1</sup>). The patterns of microbial emergence were indicative of bypass flow and led the authors to stress the need to consider the nature of soil hydraulics in addition to conventional theory of direct entry into mole drains. Critically, studies reporting on the role of macropore flow have provided strong evidence to suggest that very different conclusions can be reached regarding the efficiency of soils as bacterial filters (or particulate filters in general) depending on whether soil cores used within experiments are disturbed or intact.

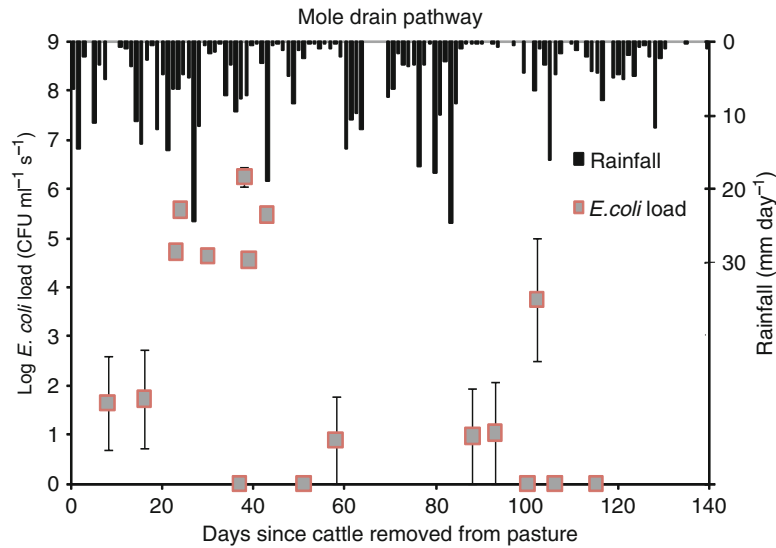
The action of disrupting macropore continuity can therefore prove effective in limiting contaminant transfer via these rapid flow conduits. For example, on arable soils there can be increased water transfer within no till soils relative to tilled soils. This is thought to be a direct influence of the increased macropore continuity within the no till soils. The action of tillage physically disrupts the soil structure and significantly reduces the extent of macropore connectivity within the soil system. Such a finding has been shown to hold true for reducing *Cryptosporidium* transfer to tile drains [80]. Tillage therefore represents a useful management approach for limiting rapid translocation of microbial pollutants (and particulate associated P) in arable soils and indicates that alternative methods of disrupting the soil structure on a no-till soil would be beneficial for reducing pathogen and FIB loss from land to water. This provides a useful management

option for arable land, but for livestock systems and grazed grasslands such approaches are obviously not possible without destruction of pasture.

The presence of subsurface drains (e.g., mole and tile drains) effectively increases subsurface connectivity provided that the age and condition of the features is such that the pathway still functions efficiently as a hydrological conduit. Field drains have been reported to be a rapid route of nutrient export from agricultural land [129] and have also been shown to export *E. coli* from land to water during storm events at similar loads to that of undrained fields using replicated 1 ha plots [82]. The use of natural fluorescence as a tracer has also demonstrated the importance of drain flow in contributing toward nonpoint pollution impacts [130]. Figure 6 shows the relationship between *E. coli* load exported via a mole drain pathway on replicated 1 ha grassland plots (clay loam) in response to rainfall events following the removal of livestock (in this case four beef steers) at the end of a 6-month grazing season. So while the installation of mole and tile drains to lower the water table may be seen as advantageous in limiting the load of potential pollutants transferred rapidly by surface runoff or near surface flow pathways it must also be appreciated that contaminants may be re-routed via different subsurface pathways [82]. Figure 6 shows clearly that significant loads of *E. coli* can be exported from drained plots, even after livestock have been removed, and that these bacteria must have transferred through the soil profile, probably via fissures and cracks in the clay soil to reach drain pathways.

### Slow Pathways of Contaminant Transfer

Much research has focused on the infiltration and vertical transfer of pathogens, FIB, and nutrients in soil leachate, and for particulate contaminants there has been considerable assessment of FIB and pathogen movement in relation to colloid filtration theory (e.g., [49, 67, 130]). Vertical displacement of microbes and nutrients through the soil profile has been demonstrated in a variety of soil column experiments (see Table 1). For bacterial transfer, the moisture content of the soil determines cell movement because continuous water films allow for transfer that is essentially limited to the aqueous phase and the solid-liquid



**Pathogen and Nutrient Transfer Through and Across Agricultural Soils. Figure 6**

*Escherichia coli* loads exported from mole and tile drains of grassland plots in relation to daily rainfall. X-axis represents day of sampling since cattle removed. Error bars represent 1 S.E. of logarithmic mean (Modified from Oliver et al. [82])

interface. Thus, it has been proposed that appreciable bacterial movement in soil can only occur if there are enough water filled pores of the required diameter to enable a continuous pathway [131]. Nitrate is undoubtedly the main form of N leached from agricultural systems, and greater concentrations will be detected in leachate from soils where N inputs on the land surface produce  $\text{NO}_3^-$  in excess of amounts required by plant uptake. The soluble and highly mobile nature of  $\text{NO}_3^-$  means that those soils in climatic zones typical of wet weather regimes will deliver potentially high concentrations of  $\text{NO}_3^-$ , though because of dilution the concentrations detected will be less.

Under conditions that promote water movement through soil pores, the carriage and translocation of microbes within the drainage water through the soil structure will depend on sieving effects imposed by pore openings. Under conditions of limited pore clogging, bacterial transfer is possible to considerable depths below the soil surface layers following the initial application of a surface applied manure source [132]. However, the transfer of microbes and particulates through the soil architecture can lead to pore clogging that may subsequently restrict such contaminant

transfer by the physical blocking of a pore entrance. Bioclogging of the pore network is clearly a function of the particle size of the porous medium and the diameter of the microbial consortia that transfer through the soil system. Bacteria, protozoa, and particulate phosphorus suspended in flow can be effectively strained by the soil matrix and then accumulate within soil passages when pore openings are too small to permit their continued transfer. These pores then become immobile regions that may exist in the filter matrix in the form of ineffective micropores, resulting in the trapping of microbes and sediments in dead end pores. Recent work by Donnison and Ross [58] investigated the effect of soil type on transfer of zoonotic bacteria to rural streams using intact cores and turfs of a gley soil and a sandy loam. Farm dairy effluent, containing laboratory grown *E. coli* O157:H7 and *Campylobacter*, was added to cores and turfs that were stored at 10°C. It was found that the relative timing of application of zoonotic bacteria to soil and that of subsequent rainfall determine whether significant numbers of these bacteria are likely to reach streams. This timing varied for different soil types. For the sandy loam soil there was little transport of cells to drainage after 14 days but for the gley soil there was little change in the proportion of

surviving *E. coli* O157:H7 that was washed out over 35 days, and it was found that *E. coli* O157 could be mobilized relatively easily from this soil. Donnison and Ross [58] highlighted that although far fewer bacteria are washed out than are retained, even immediately after application, the actual numbers can be substantial and could contribute to waterborne disease.

Under conditions of low rainfall, the soil may act as a more efficient filter matrix; this is because the slow percolation of FIB, pathogens, and P into the soil associated with light rainfall will increase microbial and P exposure time to soil surfaces in contrast to rapid water flows, and allow for an increased potential of contact with soil. This does not follow for  $\text{NO}_3^-$ , which does not interact or associate with the surrounding soil. In well-structured soils N will be retained to a high degree when protected within the bulk of the soil and will therefore only transfer through the soil profile with the slow percolation of mobile water. For microbes and P, implications of forming a cell-particle or P-particle composite may have impact once delivered to surface waters. The significant proportion of *E. coli*, for example, that remains unattached or associated with particles  $<2 \mu\text{m}$  are likely to remain in the water column and cause potential contamination problems further downstream. In contrast, the proportion of *E. coli* associated with larger soil particles are more likely to sediment out into the stream bed and therefore pose a delayed threat to water quality upon their resuspension. Further affinity studies of FIB are required, using different soil types, to test for the effects of an array of environmental variables such as pH, soil mineralogy, and temperature on the affinity of *E. coli* for soil.

Interesting observations have been made regarding the effect of cattle manure and slurry application on the percolation of the pathogens *E. coli* O157:H7 and *S. enterica* serovar Typhimurium. In the study of Semenov et al. [133] a greater number of cells were found to percolate to greater soil depths after slurry application compared to cells applied via manure application. Such results suggest that surface application of solid manures rather than their liquid counterparts may decrease the risk of contamination of groundwater supplies. Phosphate sorption (and release) in soils has also been investigated in relation to fertilizer sources.

Such sorption is affected by reactions that take place at the solution-soil surface interface and recent work has demonstrated that phosphate binding strength can be up to 50% less in manured soils than in soils fertilized with inorganic triple superphosphate [134]. Such findings imply manure applications to some soil types result in an increased net negative surface charge and therefore a reduced soil phosphate adsorption capacity leading to increased losses from the soil system. These slower vertical transfers may eventually allow for connection with groundwater supplies, which are covered in the next section of this chapter.

### The Role of Groundwater

Microbial and nutrient transport to surface water can occur by deposition of manure directly in the water or by wash-off in surface and subsurface runoff. However, transport to groundwater is a somewhat lesser concern in that it requires that the contaminant of study move through soil and bedrock to reach the water table. Nevertheless, groundwater systems are the predominant reservoir and strategic reserve of global freshwater storage at ca. 30% of the global water total and 98% of freshwater in liquid form [22], and protection of the microbial and chemical quality of this precious freshwater resource is therefore essential to safeguard human health and maintain sustainable freshwater resources. While a number of contaminants may interact with groundwater supplies, it is generally accepted that the most widely studied contaminant of groundwater supplies is  $\text{NO}_3^-$ . However, it cannot be disputed that the groundwater pollutants that most concern human health are microbiological, causing disease and sometimes death [135].

Manure application to land is regarded as the primary source of bacterial groundwater contamination in agricultural settings [136]. However, most pathogens are suspected to have life cycles far shorter (i.e., high die-off or inactivation rates outside of the animal gut) than typical groundwater travel times, except in karstic aquifers because of their dual porosity or where the source is very close to the point of water abstraction. Even so, continued investigation on microbial persistence within a range of agricultural environments is needed, and published persistence profiles of robust

pathogens such as *C. parvum* propose that this pathogen can remain viable well in excess of hundreds of days [17], suggesting potential implications for groundwater quality upon emergence. The larger size of this protozoan relative to bacterial cells means it is more likely to be filtered by the overlying soil, but direct wash-in of such pathogens into private water supplies is not unheard of and can cause significant human illness. Of course, in addition to nitrates, bacteria, and protozoa, there remain other contaminants of groundwater, their presence dictated by their mobility and reactivity with the aquifer matrix and its overlying soil [135].

A recent case study of contamination of groundwater with microbial pollutants is illustrated by a small dryland watershed in central Chile [137]. This study suggested that concentrations of FIB were temporally dynamic with levels varying between seasons with higher concentrations in winter. As discussed in earlier sections of this chapter, causes of contamination could be linked to the easy access of domestic animals to the water source (i.e., wells in this case) and to the permeable well casing material. Local precipitation runoff would also have a direct influence on the bacterial concentrations found in such wells, with seasonality influencing runoff volumes and climatic characteristics important for driving transfer. Undoubtedly, the climatic conditions typical of winter can also favor survival of pathogens and FIB, with cooler temperatures often considered advantageous for lengthy persistence of these microbes. Studies of *Cryptosporidium* through undisturbed, macroporous karst soil have demonstrated that leaching is an important mechanism of oocyst transport in karst soils but have also provided evidence for the significant role played by macropores (detailed earlier in this chapter) in the soil physical structure. Thus oocysts leaching from soils into the epikarst could accumulate and remain viable for months until hydrological conditions are right for flushing the oocysts into the conduit flow system [59].

Studies have also shown that high concentrations of nitrate in groundwater samples do not relate well to concentrations of bacterial contaminants. A Vietnamese study took measurements of  $\text{NO}_3^-$  and FIB in groundwater samples taken from both dug wells and bores and found that a significant number (18%)

of samples had  $\text{NO}_3^-$  concentrations in excess of the WHO Guideline value for drinking water of  $50 \text{ mg L}^{-1}$  ( $11.3 \text{ mg L}^{-1}$  as nitrate-nitrogen). High concentrations of FIB were found in many of the dug wells and even in the deeper drilled bores, but there was no correlation between  $\text{NO}_3^-$  concentration and bacterial content [138], probably reflecting the different transfer pathways and mobilities of these contaminant typologies through and across the surrounding landscape. Attempts to curb groundwater contamination from nitrate and FIB include novel approaches to mitigation and management such as tree planting within agricultural systems to allow for a more efficient use of resources, since the rooting system of the trees captures nutrients that are not captured by crops, for example. In particular, intercropping systems have been shown to offer potential mitigating effects on *E. coli* movement to the groundwater [92].

In contrast to  $\text{NO}_3^-$ , phosphate is mostly immobile in the subsurface, with the flux of P to groundwater being controlled by the degree of attenuation in soil. It has therefore generally been viewed as presenting minimal threat to groundwater quality. However, recent contrary views have been published contesting the long-held belief that adsorption and metal complex formation retain the majority of potentially mobile phosphorus. The relative contributions of potential sources for these elevated concentrations of P in groundwater are currently unclear but there is evidence to suggest that they are probably partly anthropogenic. Recent findings therefore suggest that groundwater P concentrations are such that they may be a more important contributor to surface water phosphorus, contrary to prior thinking [139]. Fracture flow is critically important because it allows turbulent flow to transport colloidal- and particulate-attached P to groundwater via a pathway where there is low rock–water interaction (see earlier discussion of macropore flow). Similar to concepts governing high levels of emergence of contaminants via macropore flow, the connectivity between fractures rather than fracture presence per se provide major controls on water residence time distribution from the surface system to the receiving surface water body.

Groundwater–surface water interactions are a final point to note for the upwelling of nutrient sources into

streams and rivers. Spatial and temporal distributions of  $\text{NO}_3^-$  can be observed along the upwelling flow path from groundwater to surface water along stream reaches. A UK study on a Cumbrian stream reach identified that upwelling flows dominated the exchange between groundwater and surface water throughout a period of investigation. In particular, for this stream reach,  $\text{NO}_3^-$  concentrations along upwelling flow paths appeared to follow two opposite trends, with both decreasing and increasing nitrate concentrations being observed at different points in the experimental reach. The magnitude of variation in  $\text{NO}_3^-$  concentration along the upwelling flow path to the streambed was thought to be governed by the sediment structure and characteristics in the two contrasting field sites [140]. Research studies detailing microbial pollution linked to groundwater–surface water interactions are scarce, though it is an interesting aspect for future consideration. Under baseflow conditions, the general assumption is that no resuspension of FIB occurs under steady state flow conditions, and only during storm event flows do stream sediments become resuspended in the overlying water column [141]. However, there may be potential for uncontaminated (FIB-free) groundwater to become contaminated as it travels through upwelling flow paths of the hyporheic zone and the sediment associated store of FIB, possibly adding to the baseflow load of FIB. This is clearly an underexplored research agenda for FIB at current time.

### Future Directions

Continued research conducted across a range of scales to understand better the transfer of pathogens and chemicals through and over soils is paramount. Different scales of study investigating mechanistic processes through to catchment scale real-world responses of landscape systems all serve key roles in enhancing understanding of pollutant dynamics in the soil-water continuum. Ultimately, management of livestock and their manure must be undertaken with a view to ensure the sustainability of key ecosystem services, such as the provision of clean and safe recreational and drinking water [142] and targeting of mitigation options to protect watercourses needs to be grounded on sound science. The database of such scientific results

continues to grow and provides an expanding (though this is immature for FIB and pathogens relative to P and N) body of empirical science that can form the “evidence base” for good regulatory practice [4]. A growing number of initiatives continue to emerge that advocate the drive for integrated catchment management and the need for greater stakeholder engagement to help reduce contaminant transfers. For successful mitigation of pollutant transfers through and across soil that reduce impact on water quality there is a clear need to tailor mitigation options to reap the most “bang for the buck,” and therefore management interventions and changes must be considered in terms of both cost and efficiency [143]. Making informed decisions at the field scale is therefore crucial because agricultural land is heterogeneous, and inherent spatial variability in soils and hydrological flow pathways influences the loss of pollutants from land to water, often at subfield scales [144]. Clearly, land management options such as manure incorporation into soil and reduced application rates to pasture can help avoid scenarios likely to enhance transport and serve as viable approaches to minimize nonpoint-source contamination while ensuring the least public health risk [64]. However, mitigation strategies are often much more complex to implement in practice and accommodate site specific constraints. The lag time between implementation of management practices on the land and water quality response is an unfortunate reality in watershed management and adds an extra layer of complexity to matters [36].

While efforts continue in an attempt to limit contaminant transfers through and across soils, there are parallel challenges for the research and policy community. It is critical that soil and water scientists make progress on quantifying losses from (and delivery via) spatially and temporally dynamic pathways rather than only appreciating the existence of such pathways for facilitating pollutant transfers. In particular, quantifying exports of pathogens remains elusive at farm and catchment scales. However, future research agendas for pathogen transport through soil systems must accommodate the need to characterize microbial behavior across a broad range of environmentally relevant conditions. Critically, the transfer pathways of both pathogens and chemicals will be impacted by the likely

changes in climatic conditions [43], which may also impact indirectly on contaminant transfers because of changes to transport and movement of animals, intensity of livestock farming, and habitat change [145, 146]. Flooding too represents a transfer mechanism by which pathogens, FIB, and nutrients can be catastrophically exported from agricultural land in considerable numbers and result in much greater exposure of the human population to pollutants [3, 147]. Thus flooding represents an example of a very different route of transfer, which at present remains poorly understood [148] but which is sure to warrant much attention as an exposure pathway of increasing significance.

What this chapter has attempted to show, in part, is that various pollutant typologies (soluble, particulate, inert, biological etc.) behave and respond to environmental drivers and conditions in different ways. Pathways through and across the soil are used to a different extent by different pollutant types and are influenced by spatial and temporal characteristics of rainfall and soil types. The warning message for mitigation options implemented to reduce pollutant transfer through soil systems is that different strategies will impact on pollutants in different ways. A mitigation strategy designed to impede particulate transfer may, in fact, enhance the movement of soluble contaminants, or even impact on greenhouse gas emissions. The need for holistic understanding of multi-pollutants is now essential to reflect the potential for such “pollution swapping” risks [149, 150]. In turn, approaches for modeling agricultural systems and the flows and transfers of associated contaminants within the soil-water continuum now require frameworks that can explore the effectiveness of nonpoint pollutant mitigation options for multiple pollutants (including those designed for pathogens). Undoubtedly, such tools are high on the agenda for policy-makers who are required to identify which mitigation options are likely to reduce the target pollutant without increasing impact from others [151]. Indeed, integrated water and agricultural management is needed so that decision makers can recognize interdependencies in environmental systems and prioritize management responsibilities for optimal protection of environmental resources [152, 153].

Finally, new technologies continue to push forward the boundaries of science. Higher resolution and

continuous sampling is highlighting trends in diffuse pollutants previously undetected and substantiating our knowledge of catchment response to environmental drivers. For detection of microbial pollutants and associated transfer routes through the environment, the future is likely to embrace molecular approaches. Studies are beginning to emerge that investigate molecular versus culture-based approaches for the detection of FIB, and such initiatives are key in examining relationships between quantitative PCR (qPCR) and culture-based FIB counts in an attempt to optimize and standardize methods for current indicators. However, currently, very few studies have attempted to quantify the uncertainty in quantitative qPCR data, though this is likely to be rectified over the coming years. Indeed, such tools are continuing to be investigated, developed, and tested on “end-point receptor” waters (e.g., coastal waters) (e.g., [154]). With continued evaluation of these approaches against standard culture-based techniques (for bacterial studies), there is much potential for these methods to be developed further to determine potential sources and pathways of pollution linked to both surface and groundwater contamination.

## Bibliography

### Primary Literature

1. Haygarth PM, Ritz K (2009) The future of soils and land use in the UK: soil systems for the provision of land-based ecosystem services. *Land Use Policy* 26S:5187–5197
2. Pepper IL, Gerba CP, Newby DT, Rice CW (2009) Soil: a public health threat or savior? *Crit Rev Environ Sci Technol* 39:416–432
3. Bridge JW, Oliver DM, Chadwick D, Godfray C, Heathwaite AL, Kay D, Maheswaran R, McGonigle D, Nichols G, Pickup R, Porter J, Wastling J, Banwart SA (under review). Minimising the waterborne disease burden in affluent nations: a key role for environmental sciences. *Bull World Health Organ*
4. Kay D, Crowther J, Fewtrell L, Francis CA, Hopkins M, Kay C, McDonald AT, Stapleton CM, Watkins J, Wyer MD (2008) Quantification and control of microbial pollution from agriculture: a new policy challenge? *Environ Sci Policy* 11:171–184
5. Kay D, Edwards AC, Ferrier RC, Francis C, Kay C, Rushby L, Watkins J, McDonald AT, Wyer M, Crowther J, Wilkinson J (2007) Catchment microbial dynamics: the emergence of a research agenda. *Prog Phys Geogr* 31:59–76
6. USEPA (2010) National summary of impaired waters and TMDL information. U.S. Environmental Protection Agency,

- Washington. Available at: [http://iaspub.epa.gov/waters10/attains\\_nation\\_cy.control?p\\_report\\_type=T](http://iaspub.epa.gov/waters10/attains_nation_cy.control?p_report_type=T). Accessed 4 Feb 2010
7. Hunter PR (1992) Cyanobacteria and human health. *J Med Microbiol* 36:301–302
  8. Pang L, Nowostawska U, Ryan JN, Williamson WM, Walshe G, Hunter KA (2009) Modifying the surface charge of pathogen-sized microspheres for studying pathogen transport in groundwater. *J Environ Qual* 38:2210–2217
  9. Infascelli R, Pelorosso R, Boccia L (2009) Spatial assessment of animal manure spreading and groundwater nitrate pollution. *Geospat Health* 4:27–38
  10. Bartram J, Thyssen N, Gowers A, Pond K, Lack T (2002) Water and health in Europe. A joint report from the European Environment Agency and the WHO regional office for Europe. WHO Regional Publications, Copenhagen
  11. Stewart I, Webb PM, Schluter PJ, Shaw GR (2006) Recreational and occupational field exposure to freshwater cyanobacteria – a review of anecdotal and case reports, epidemiological studies and the challenges for epidemiologic assessment. *Environ Health* 5:6
  12. Topp E, Scott A, Lapen DR, Lyautey AR, Duriez P (2009) Livestock waste treatment systems for reducing environmental exposure to hazardous enteric pathogens: some considerations. *Bioresour Technol* 100:5395–5398
  13. Monaghan RM, Carey PL, Wilcock RJ, Drewry JJ, Houlbrooke DJ, Quinn JM, Thorrold BS (2009) Linkages between land management activities and stream water quality in a border dyke-irrigated pastoral catchment. *Agric Ecosyst Environ* 129:201–211
  14. McKergow LA, Davies-Colley RJ (2009) Stormflow dynamics and loads of *Escherichia coli* in a large mixed land use catchment. *Hydrol Process* 24:276–289
  15. Kay D, Anthony S, Crowther J, Chambers BJ, Nicholson FA, Chadwick D, Stapleton CM, Wyer MD (2010) Microbial water pollution: a screening tool for initial catchment-scale assessment and source apportionment. *Sci Total Environ* (in press)
  16. Arinaminpathy N, McLean AR, Godfray HCJ (2009) Future UK land use policy and the risk of infectious disease in humans, livestock and wild animals. *Land use Policy* 26:5124–5133
  17. Oliver DM, Clegg CD, Haygarth PM, Heathwaite AL (2005a) Assessing the potential for pathogen transfer from grassland soils to surface waters. *Adv Agron* 85:125–180
  18. Sobsey MD, Pillai SD (2009) Where future emerging pathogens will come from and what approaches can be used to find them, besides VFARs. *J Water Health* 7:575–593
  19. St.-Pierre K, Levesque S, Frost E, Carrier N, Arbeit RD, Michaud S (2009) Thermotolerant coliforms are not a good surrogate for *Campylobacter* spp. in environmental waters. *Appl Environ Microbiol* 75:6736–6744
  20. Wilkes G, Edge T, Gannon V, Jokinen C, Lyautey E, Medeiros D, Neumann N, Ruecker N, Topp E, Lapen D (2009) Seasonal relationships among indicator bacteria, pathogenic bacteria, *Cryptosporidium* oocysts, *Giardia* cysts and hydrological indices for surface waters within an agricultural landscape. *Water Res* 43:2209–2223
  21. Conley DJ, Paerl HW, Howarth HW, Boesch DF, Seitzinger SP, Havens KE, Lancelot C, Likens GE (2009) Controlling eutrophication: nitrogen and phosphorus. *Science* 323:1014–1015
  22. Heathwaite AL (2010) Multiple stressors on water availability at global to catchment scales: understanding human impact on nutrient cycles to protect water quality and water availability in the long term. *Freshw Biol* 55:241–257
  23. Edmeades DC (2003) The long-term effects of manures and fertilizers on soil productivity and quality: a review. *Nutr Cycl Agroecosyst* 66:165–180
  24. Haygarth PM, Heathwaite AL, Jarvis SC, Harrod TR (2000) Hydrological factors for phosphorus transfer from agricultural soils. *Adv Agron* 69:153–178
  25. Muirhead RW (2009) Soil and fecal material reservoirs of *Escherichia coli* in a grazed pasture. *NZ J Agric Res* 52:1–8
  26. Oliver DM, Page T, Heathwaite AL, Haygarth PM (2010a) Reshaping models of *E. coli* population dynamics in livestock feces: increased bacterial risk to humans? *Environ Int* 36:1–7
  27. Chadwick DR, Fish RD, Oliver DM, Heathwaite AL, Hodgson CJ, Winter M (2008) Management of livestock and their manure to reduce the risk of microbial transfers to water – the case for an interdisciplinary approach. *Trends Food Sci Technol* 19:240–247
  28. Oliver DM, Heathwaite AL, Fish RD, Chadwick DR, Hodgson CJ, Winter M, Butler AJ (2009) Scale appropriate modeling of diffuse microbial pollution from agriculture. *Prog Phys Geogr* 33:358–377
  29. Unc A, Goss MJ (2006) Culturable *Escherichia coli* in soil mixed with two types of manure. *Soil Sci Soc Am J* 70:763–769
  30. Oliver DM, Page T, Hodgson CJ, Heathwaite AL, Chadwick DR, Fish RD, Winter M (2010b) Development and testing of a risk indexing framework to determine fields-scale critical source areas of faecal bacteria on grassland. *Environ Modell Softw* 25:503–512
  31. Oliver DM, Haygarth PM, Clegg CD, Heathwaite AL (2006) - Differential *E. coli* die off patterns associated with agricultural matrices. *Environ Sci Technol* 40:5710–5716
  32. Hodgson CJ, Bulmer N, Chadwick DR, Oliver DM, Heathwaite AL, Fish RD, Winter M (2009) Establishing relative release kinetics of fecal indicator organisms from different fecal matrices. *Lett Appl Microbiol* 49:124–130
  33. Brennan FP, O’Flaherty V, Kramers G, Grant J, Richards KG (2010) Long-term persistence and leaching of *E. coli* in temperate maritime soils. *Appl Environ Microbiol* 76(5):1449–1455
  34. Lyautey E, Zexun L, Lapen DR, Wilkes G, Scott A, Berkens T, Edge TA, Topp E (2010) Distribution and diversity of *Escherichia coli* populations in the South Nation River Drainage Basin, Eastern Ontario, Canada. *Appl Environ Microbiol* 76(5):1486–1496
  35. Nautiyal CS, Rehman A, Chauhan PS (2010) Environmental *Escherichia coli* occur as natural plant growth promoting soil bacterium. *Arch Microbiol* 192:185–193

36. Meals DW, Dressing SA, Davenport TE (2009) Lag time in water quality response to best management practices: a review. *J Environ Qual* 39:85–96
37. Preedy N, McTiernan K, Matthews R, Heathwaite AL, Haygarth PM (2001) Rapid incidental phosphorus transfers from grassland. *J Environ Qual* 30:2105–2122
38. Hutchison ML, Walters LD, Moore A, Crookes KM, Avery SM (2004) Effect of length of time before incorporation on survival of pathogenic bacteria present in livestock wastes applied to agricultural soil. *Appl Environ Microbiol* 70:5111–5118
39. Oliver DM, Heathwaite AL, Hodgson CJ, Chadwick DR (2007a) Mitigation and current management attempts to limit pathogen survival and movement within farmed grasslands. *Adv Agron* 93:95–152
40. McGechan MB, Topp CFE (2004) Modeling environmental impacts of deposition of excreted nitrogen by dairy cows. *Agric Ecosyst Environ* 103:149–164
41. Aitken MN (2003) Impact of agricultural practices and river catchment characteristics on river and bathing water quality. *Water Sci Technol* 48:217–224
42. Deasy C, Brazier RE, Heathwaite AL, Hodgkinson R (2009) Pathways of runoff and sediment transfer in small agricultural catchments. *Hydrol Process* 23:1349–1358
43. Boxall ABA, Hardy A, Beulke S, Boucard T, Burgin L, Falloon PD, Haygarth PM, Hutchinson T, Kovats RS, Leonardi G, Levy LS, Nichols G, Parsons SA, Potts L, Stone D, Topp E, Turley DB, Walsh K, Wellington EMH, Williams RJ (2009) Impacts of climate change on indirect human exposure to pathogens and chemicals from agriculture. *Environ Health Perspect* 117: 508–514
44. Garbrecht K, Fox GA, Guzman JA, Alexander D (2009) *E. coli* transport through soil columns: implications for bioretention cell removal efficiency. *Trans ASABE* 52:481–486
45. Armstrong A, Quinton JN (2010) Variability of interrill erosion at low slopes. *Earth Surf Process Landforms*
46. Lane SN, Brookes CJ, Kirkby AJ, Holden J (2004) A network-indexbased version of TOPMODEL for use with high-resolution digital topographic data. *Hydrol Process* 18:191–201
47. Reaney SM, Lane SN, Heathwaite AL, Dugdale LJ (2010) Risk-based modeling of diffuse land use impacts from rural landscapes upon salmonid fry abundance. *Ecol Model* (in press)
48. Yaron B, Dror I, Berkowitz B (2010) Contaminant geochemistry – a new perspective. *Naturwissenschaften* 97:1–17
49. Bridge JW, Banwart SA, Heathwaite AL (2006) Noninvasive quantitative measurement of colloid transport in mesoscale porous media using time lapse fluorescence imaging. *Environ Sci Technol* 40:5930–5936
50. Zonia L, Bray D (2009) Swimming patterns and dynamics of simulated *Escherichia coli* bacteria. *J R Soc Interface* 6:1035–1046
51. Tu YH, Shimizu TS, Berg HC (2008) Modelling the chemotactic response of *Escherichia coli* to time-varying stimuli. *Proc Natl Acad Sci USA* 105:14855–14860
52. Sharpley AN, Kleinman PJA, Heathwaite AL, Bburek WJ, Folmar GJ, Schmidt JR (2008) Phosphorus loss from an agricultural watershed as a function of storm size. *J Environ Qual* 37:362–368
53. Granger SJ, Hawkins JMB, Bol R, White SM, nadan P, Old G, Bilotta GS, Brazier RE, Macleod CJA, Haygarth PM (2010) High temporal resolution monitoring of multiple pollutant responses in drainage from an intensively managed grassland catchment caused by a summer storm. *Water Air Soil Pollut* 205:377–393
54. Artz RRE, Townend J, Brown K, Towers W, Killham K (2005) Soil macropores and compaction control the leaching potential of *Escherichia coli* O157:H7. *Environ Microbiol* 7:241–248
55. Mawdsley JL, Brooks AE, Merry RJ (1996) Movement of the protozoan pathogen *Cryptosporidium parvum* through three contrasting soil types. *Biol Fertil Soils* 21:30–36
56. Matula J (2009) Possible phosphorus losses from the top layer of agricultural soils by rainfall simulations in relation to multi-nutrient soil tests. *Plant Soil Environ* 55:511–518
57. Entry JA, Sojka RE, Hicks BJ (2010) Matrix-based fertilizers reduce nutrient and bacterial leaching after manure application in a greenhouse column study. *J Environ Qual* 39:384–392
58. Donnison A, Ross C (2009) Survival and retention of *Escherichia coli* O157:H7 and *Campylobacter* in contrasting soils from the Toenepi catchment. *NZ J Agric Res* 52:133–144
59. Boyer DG, Kuczynska E, Fayer R (2009) Transport, fate, and infectivity of *Cryptosporidium parvum* oocysts released from manure and leached through macroporous soil. *Environ Geol* 58:1011–1019
60. Tarkalson DD, Leytem AB (2009) Phosphorus mobility in soil columns treated with dairy manures and commercial fertilizer. *Soil Sci* 174:73–80
61. Miller JJ, Beasley BW, Chanasyk DS, Larney FJ, Olson BM (2008) Short-term nitrogen leaching potential of fresh and composted beef cattle manure applied to disturbed soil cores. *Compost Sci Util* 16:12–19
62. Harter T, Atwill ER, Hou L, Karle BM, Tate KW (2008) Developing risk models of *Cryptosporidium* transport in soils from vegetated, tilted, soilbox experiments. *J Environ Qual* 37: 245–258
63. Brock EH, Ketterings QM, Kleinman PJA (2007) Measuring and predicting the phosphorus sorption capacity of manure-amended soils. *Soil Sci* 172:266–278
64. Horswell J, Hewitt J, Prosser J, Van Schaik A, Croucher D, Macdonald C, Burford P, Susarla P, Bickers P, Speir T (2010) Mobility and survival of *Salmonella Typhimurium* and human adenovirus from spiked sewage sludge applied to soil columns. *J Appl Microbiol* 108:104–114
65. Rosa BA, Yim MI, Burdenuk L, Kjartanson B, Leung KT (2010) The transport of *Escherichia coli* through freeze-fractured clay soil. *Water Air Soil Pollut* 210:243–254
66. Bech TB, Johnsen K, Dalsgaard A, Laegdsmand M, Jacobsen OH, Jacobsen CS (2010) Transport and distribution of *Salmonella enterica* serovar Typhimurium in loamy and sandy soil monoliths with applied liquid manure. *Appl Environ Microbiol* 76:710–714



67. Turner BL, Haygarth PM (2000) Phosphorus forms and concentrations in leachate under four grassland soil types. *Soil Sci Soc Am J* 64:1090–1099
68. Di HJ, Cameron KC, Shen JP, He JZ, Winefield JS (2009) A lysimeter study of nitrate leaching from grazed grassland as affected by a nitrification inhibitor, dicyandiamide, and relationships with ammonia oxidizing bacteria and archaea. *Soil Use Manage* 25:454–451
69. Guzman JA, Fox GA, Malone RW, Kanwar RS (2009) *Escherichia coli* transport from surface-applied manure to subsurface drains through artificial biopores. *J Environ Qual* 38: 2412–2421
70. Muirhead RW, Collins RP, Bremer PJ (2006) Interaction of *Escherichia coli* and soil particles in runoff. *Appl Environ Microbiol* 72:3406–3411
71. Ferguson CM, Davies CM, Kaucner C, Krogh M, Rodehutsors J, Deere DA, Ashbolt NJ (2007) Field scale quantification of microbial transport from bovine feces under simulated rainfall events. *J Water Health* 5:83–95
72. Abu-Ashour J, Lee H (2000) Transport of bacteria on sloping soil surfaces by runoff. *Environ Toxicol* 15:149–153
73. Tunney H, Kurz I, Bourke D, O'Reilly C (2007) RTDI programme 2000–2006 eutrophication from agricultural sources: the impact of the grazing animal on phosphorous loss from grazed pasture. Teagasc, Wexford
74. Quinton JN, Catt JA, Hess TM (2001) The selective removal of phosphorus from soil: is event size important? *J Environ Qual* 30:538–545
75. Soupir ML, Mostaghimi S, Yagow ER, Hagedorn C, Vaughan DH (2006) Transport of fecal bacteria from poultry litter and cattle manures applied to pastureland. *Water Air Soil Pollut* 169:125–136
76. Withers PJA, Hodgkinson RA, Bates A, Withers CL (2007) Soil cultivation effects on sediment and phosphorus mobilization in surface runoff from three contrasting soil types in England. *Soil Tillage Res* 93: 438–451
77. Mishra A, Benham BL, Mostaghimi S (2008) Bacterial transport from agricultural lands fertilized with animal manure. *Water Air Soil Pollut* 189:127–134
78. Heathwaite AL, Haygarth PM, Matthews R, Preedy N, Butler P (2005) Evaluating colloidal phosphorus delivery to surface waters from non-point agricultural sources. *J Environ Qual* 34:287–298
79. Thiagarajan A, Gordon R, Madani A, Stratton GW (2007) Discharge of *Escherichia coli* from agricultural surface and subsurface drainage water: tillage effects. *Water Air Soil Pollut* 182:3–12
80. Ramirez NE, Wang P, Lejeune J, Shipitalo MJ, Ward LA, Sreevatsan S, Dick WA (2009) Effect of tillage and rainfall on transport of manure-applied *Cryptosporidium parvum* oocysts through soil. *J Environ Qual* 38:2394–2401
81. Haygarth PM, Hepworth L, Jarvis SC (1998) Forms of phosphorus transfer in hydrological pathways from soil under grazed grassland. *Eur J Soil Sci* 49:65–72
82. Oliver DM, Heathwaite AL, Haygarth PM, Clegg CD (2005) Transfer of *Escherichia coli* to water from drained and undrained grassland after grazing. *J Environ Qual* 34:918–925
83. Alfaro M, Salazar F, Iraira S, Teuber N, Villarroel D, Ramirez L (2008) Nitrogen, phosphorus and potassium losses in a grazing system with different stocking rates in a volcanic soil. *Chil J Agric Res* 68:146–155
84. Kouznetsov MY, Roodsari R, Pachepsky YA, Shelton DR, Sadeghi AM, Shirmohammadi A, Starr JL (2007) Modeling manure-borne bromide and fecal coliform transport with runoff and infiltration at a hillslope. *J Environ Manage* 84:336–346
85. Olson BM, Bennett DR, McKenzie RH, Ormann TD, Atkins RP (2009) Nitrate leaching in two irrigated soils with different rates of cattle manure. *J Environ Qual* 38:2218–2228
86. Collins R, Elliot S, Adams R (2005) Overland flow delivery of fecal bacteria to a headwater pastoral stream. *J Appl Microbiol* 99:126–132
87. Kay D, Aitken M, Crowther J, Dickinson I, Edwards AC, Francis C, Hopkins M, Jeffrey W, Kay C, McDonald AT, McDonald D, Stapleton CM, Watkins J, Wilkinson J, Wyer M (2007) Reducing fluxes of fecal indicator compliance parameters to bathing waters from non-point agricultural sources: the Brighouse Bay study. *Scotland Environ Pollut* 147:138–149
88. Deasy C, Heathwaite AL, Brazier RE (2008) A field methodology for quantifying phosphorus transfer and delivery to streams in first order agricultural catchments. *J Hydrol* 350:329–338
89. Botter G, Milan E, Bertuzzo E, Zanardo S, Marani M, Rinaldo A (2009) Inferences from catchment-scale tracer circulation experiments. *J Hydrol* 369:368–380
90. Close M, Dann R, Ball A, Pirie R, Savill M, Smith Z (2008) Microbial groundwater quality and its health implications for a border-strip irrigated dairy farm catchment, South Island, New Zealand. *J Water Health* 6:83–98
91. Heathwaite AL, Johnes PJ (1996) Contribution of nitrogen species and phosphorus fractions to stream water quality in agricultural catchments. *Hydrol Process* 10:971–983
92. Dougherty MC, Thevathasan NV, Gordon AM, Lee H, Kort J (2009) Nitrate and *Escherichia coli* NAR analysis in tile drain effluent from a mixed tree intercrop and monocrop system. *Agric Ecosyst Environ* 131:77–84
93. Davies-Colley R, Lydiard E, Nagels J (2008) Stormflow-dominated loads of fecal pollution from an intensively dairy-farmed catchment. *Water Sci Technol* 57:1519–1523
94. Rothwell JJ, Dise NB, Taylor KG, Allott TEH, Scholefield P, Davies H, Neal C (2010) A spatial and seasonal assessment of river water chemistry across North West England. *Sci Total Environ* 408:841–855
95. Beven K, Heathwaite AL, Haygarth PM, Walling D, Brazier R, Withers P (2005) On the concept of delivery of sediment and nutrients to stream channels. *Hydrol Process* 19:551–556
96. Haygarth PM, Condon LM, Heathwaite AL, Turner BL, Harris GP (2005) The phosphorus transfer continuum: linking source to impact with an interdisciplinary and multi-scaled approach. *Sci Total Environ* 344:5–14

97. Armstrong A, Quinton JN (2009) Pumped rainfall simulators: the impact of rain pulses on sediment concentration and size. *Earth Surf Process Land* 34:1310–1314
98. Mankin KR, Wang L, Hutchinson SL, Marchin GL (2007) *Escherichia coli* sorption to sand and silt loam soil. *Trans ASABE* 50:1159–1165
99. Penn CJ, Mullins GL, Zelazny LW (2005) Mineralogy in relation to phosphorus sorption and dissolved phosphorus losses in runoff. *Soil Sci Soc Am J* 69:1532–1540
100. Drozd C, Schwartzbrod J (1996) Hydrophobic and electrostatic cell surface properties of *Cryptosporidium parvum*. *Appl Environ Microbiol* 62:1227–1232
101. Kuczynska E, Shelton DR, Pachepsky Y (2005) Effect of bovine manure on *Cryptosporidium parvum* oocyst attachment to soil. *Appl Environ Microbiol* 71:6394–6397
102. Voice TC, Weber WJ (1985) Sorbent concentration effects in liquid/solid partitioning. *Environ Sci Technol* 19:789–796
103. Oliver DM (2005) Hydrological pathways and processes of *Escherichia coli* transfer from grassland soils to surface waters. Ph.D. thesis, University of Sheffield
104. Koopmans GF, McDowell RW, Chardon WJ, Oenema O, Dolfing J (2002) Soil phosphorus quantity-intensity relationships to predict increased soil phosphorus loss to overland and subsurface flow. *Chemosphere* 48:679–687
105. Ling TY, Achberger EC, Drapcho CM, Bengtson RL (2002) Quantifying adsorption of an indicator bacteria in a soil-water system. *Trans ASAE* 45:669–674
106. Kretzschmar R, Borkovec M, Grolimund D, Elimelech M (1999) Mobile subsurface colloids and their role in contaminant transport. *Adv Agron* 66:121–193
107. Soupier ML, Mostaghimi S, Dillaha T (2010) Attachment of *Escherichia coli* and Enterococci to particles in runoff from bare soils. *J Environ Qual* (in press)
108. Oliver DM, Clegg CD, Heathwaite AL, Haygarth PM (2007b) Preferential attachment of *Escherichia coli* to different particle size fractions of an agricultural grassland soil. *Water Air Soil Pollut* 185:369–375
109. Muirhead RW, Collins RP, Bremer PJ (2006) The association of *E. coli* and soil particles in overland flow. *Water Sci Technol* 54:153–159
110. Pachepsky YA, Shelton DR (2011) *Escherichia coli* and fecal coliforms in freshwater and estuarine sediments. *Crit Rev Environ Sci Technol* (in press)
111. Sinclair A, Hebb D, Jamieson R, Gordon R, Benedict K, Fuller K, Stratton GW, Madani A (2009) Growing season surface water loading of fecal indicator organisms within a rural watershed. *Water Res* 43:1199–1206
112. Haygarth PM, Wood FL, Heathwaite AL, Butler P (2005) Phosphorus dynamics observed through increasing scales in a nested headwater-to-river channel study. *Sci Total Environ* 344:83–106
113. Jordan P, Arnscheidt J, McGrogan H, McCormick S (2007) Characterising phosphorus transfers in rural catchments using a continuous bank-side analyser. *Hydrol Earth Syst Sci* 11:372–381
114. Tyrrel SF, Quinton JN (2003) Overland flow transport of pathogens from agricultural land receiving fecal wastes. *J Appl Microbiol* 94:875–935
115. McDonald A, Kay D (1981) Enteric bacterial concentrations in reservoir feeder streams – baseflow characteristics and response to hydrograph events. *Water Res* 15:961–968
116. Heathwaite AL, Burt TP, Trudgill ST (1990) The effect of land use on nitrogen, phosphorus and suspended sediment delivery to streams in a small catchment in southwest England. In: Boardman J, Foster LDL, Dearing JA (eds) *Soil erosion on agricultural land*. Wiley, Chichester/New York/Brisbane/Toronto/Singapore, pp 161–177
117. Page T (2008) Uncertainty assessment of phosphorus risk to surface waters. Environment Agency Science Report – SC050035
118. Kirchner JW, Feng XH, Neal C (2000) Fractal stream chemistry and its implications for contaminant transport in catchments. *Nature* 403:524–527
119. Beven K, Germann P (1982) Macropores and water-flow in soils. *Water Resour Res* 18:1311–1325
120. Abu-Ashour J, Joy DM, Lee H, Whiteley HR, Zelin S (1998) Movement of bacteria in unsaturated soil columns with macropores. *Trans ASAE* 41:1043–1050
121. Geohring LD, McHugh OV, Walter MT, Steenhuis TS, Akhtar MS (2001) Phosphorus transport into subsurface drains by macropores after manure applications: implications for best manure management practices. *Soil Sci* 166:896–909
122. Allaire SE, Roulier S, Cessna AJ (2009) Quantifying preferential flow in soils: a review of different techniques. *J Hydrol* 378:179–204
123. Guber AK, Pachepsky YA, Shelton DR, Yu O (2009) Association of fecal coliforms with soil aggregates: effect of water content and bovine manure application. *Soil Sci* 174:543–548
124. Lutterodt G, Basnet M, Foppen JWA, Uhlenbrook S (2009) The effect of surface characteristics on the transport of multiple *Escherichia coli* isolates in large scale columns of quartz sand. *Water Res* 43:595–604
125. McLeod M, Aislabie J, Ryburn J, McGill A (2008) Regionalizing potential for microbial bypass flow through New Zealand soils. *J Environ Qual* 37:1959–1967
126. Aislabie J, Smith JJ, Fraser R, McLeod M (2001) Leaching of bacterial indicators of faecal contamination through four New Zealand soils. *Aust J Soil Res* 39:1397–1406
127. Paterson E, Kemp JS, Gammack SM, Fitzpatrick EA, Cresser MS, Mullins CE, Killham K (1993) Leaching of genetically-modified *Pseudomonas fluorescens* through intact soil microcosms – influence of soil type. *Biol Fertil Soils* 15:308–314
128. McLeod M, Aislabie J, Ryburn J, McGilla A, Taylor M (2003) Microbial and chemical tracer movement through two south-land soils, New Zealand. *Aust J Soil Res* 41:1163–1169
129. Heathwaite AL, Burke SP, Bolton L (2006) Field drains as a route of rapid nutrient export from agricultural land receiving biosolids. *Sci Total Environ* 365:33–46

130. Mosaddeghi MR, Mahboubi AA, Zandsalimi S, Unc A (2009) Influence of organic waste type and soil structure on the bacterial filtration rates in unsaturated intact soil columns. *J Environ Manage* 90:730–739
131. Bowen GD, Rovira AD (1999) The rhizosphere and its management to improve plant growth. *Adv Agron* 66:1–102
132. Gagliardi JV, Karns JS (2000) Leaching of *Escherichia coli* O157:H7 in diverse soils under various agricultural management practices. *Appl Environ Microbiol* 66:877–883
133. Semenov AV, van Overbeek L, van Brugger AHC (2009) Percolation and survival of *Escherichia coli* O157:H7 and *Salmonella enterica* Serovar Typhimurium in soil amended with contaminated dairy or slurry. *Appl Environ Microbiol* 75:3206–3215
134. Jiao Y, Whalen JK, Hendershot WH (2007) Phosphate sorption and release in a sandy-loam soil as influenced by fertilizer sources. *Soil Sci Soc Am J* 71:118–124
135. Lerner D, Harris B (2009) The relationship between land use and groundwater resources and quality. *Land Use Policy* 26S: S265–S273
136. Pachepsky YA, Sadeghi AM, Bradford SA, Shelton DR, Guber AK, Dao T (2006) Transport and fate of manure-borne pathogens: modeling perspective. *Agric Water Manage* 86:81–92
137. Valenzuela M, Lagos B, Claret M, Mondaca MA, Perez C, Parra O (2009) Fecal contamination of groundwater in a small rural dryland watershed in rural Chile. *Chil J Agric Res* 69:235–243
138. Cam PD, Lan NTP, Smith GD, Verma N (2008) Nitrate and bacterial contamination in well waters in Vinh Phuc province, Vietnam. *J Water Health* 6:275–279
139. Holman IP, Whelan MJ, Howden NJK, Bellamy PH, Willby NJ, Rivas-Casado M, McConvey P (2008) Phosphorus in groundwater—an overlooked contributor to eutrophication? *Hydrol Process* 22:5121–5126
140. Krause S, Heathwaite L, Binley A, Keenan P (2009) Nitrate concentration changes at the groundwater-surface water interface of a small Cumbrian river. *Hydrol Process* 23:2195–2211
141. Rehmann CR, Soupir ML (2009) Importance of interactions between the water column and the sediment for microbial concentrations in streams. *Water Res* 43:4579–4589
142. Pretty J (2008) Agricultural sustainability: concepts, principles and evidence. *Philos Trans R Soc B Biol Sci* 363: 447–465
143. Haygarth PM, ApSimon H, Betson M, Harris D, Hodgkinson R, Withers PJA (2009) Mitigating non-point phosphorus transfer from agriculture according to cost and efficiency. *J Environ Qual* 38:2012–2022
144. Hewett CJM, Quinn PF, Whitehead PG, Heathwaite AL, Flynn NJ (2004) Towards a nutrient export risk matrix approach to managing agricultural pollution at source. *Hydrol Earth Syst Sci* 8:834–845
145. Gale P, Drew T, Phipps LP, David G, Wooldridge M (2009) The effect of climate change on the occurrence and prevalence of livestock diseases in Great Britain: a review. *J Appl Microbiol* 106:1409–1423
146. Semenza JC, Menne B (2009) Climate change and infectious diseases in Europe. *Lancet Infect Dis* 9:365–375
147. Nagels JW, Davies-Colley RJ, Donnison AM, Muirhead RW (2002) Faecal contamination over flood events in a pastoral agricultural stream in New Zealand. *Water Sci Technol* 12:45–52
148. Wheeler H, Evans E (2009) Land use, water management and future flood risk. *Land Use Policy* 26S: S251–S264
149. Stevens CJ, Quinton JN (2009a) Policy implications of pollution swapping. *Phys Chem Earth* 34:589–594
150. Stevens CJ, Quinton JN (2009b) Non-point pollution swapping in arable agricultural systems. *Crit Rev Environ Sci Technol* 39:478–520
151. Cuttle SP, Macleod CJA, Chadwick DR, Scholefield D, Haygarth PM, Newell-Price P, Harris D, Shepherd MA, Chambers BJ, Humphrey R (2007) An inventory of methods to control non-point water pollution from agriculture (DWPA) user manual. Defra, London, Defra project code ES0203, 113p
152. Fish RD, Ioris AAR, Watson NM (2010) Integrating water and agricultural management: collaborative governance for a complex policy problem. *Sci Total Environ* (in press)
153. Macleod CJA, Scholefield D, Haygarth PM (2007) Integration for sustainable catchment management. *Sci Total Environ* 373:591–602
154. Fremaux B, Gritzfeld J, Boa T, Yost CK (2009) Evaluation of host specific bacteroidales 16S rRNA gene markers as complementary tool for detecting fecal pollution in a prairie watershed. *Water Res* 43:4838–4849
155. Naden PS, Old GH, Eliot-Laize C, Granger SJ, Hawkins JMB, Bol R, Haygarth P (2010) Assessment of natural fluorescence as a tracer of non-point agricultural pollution from slurry spreading on intensely-farmed grasslands. *Water Res* 44(6):1701–1712

## Books and Reviews

- Burton CH, Turner C (2003) *Manure management: treatment strategies for sustainable agriculture*, 2nd edn. Silsoe Research Institute, Bedford
- Ferguson C, Husman AMD, Altavilla N, Deere DA, Ashbolt N (2003) Fate and transport of surface water pathogens in watersheds. *Crit Rev Environ Sci Technol* 33:299–361
- Haygarth PM, Jarvis SC (2002) *Agriculture, hydrology and water quality*. CAB International, Wallingford
- McDowell RW, Houlbrooke DJ, Muirhead RW, Mueller K, Shepherd M, Cuttle SP (2008) *Grazed pastures and surface water quality*. Nova Science Publishers, Hauppauge, p 238
- Unc A, Goss MJ (2004) Transport of bacteria from manure and protection of water resources. *Appl Soil Ecol* 25:1–18
- Whitehead PG, Wilby RL, Batterbee RW, Kernan M, Wade AJ (2009) A review of the potential impacts of climate change on surface water quality. *Hydrol Sci J* 54:101–123

## PEM Fuel Cell Materials: Costs, Performance and Durability

A. DE FRANK BRUIJN<sup>1,2</sup>, GABY J. M. JANSSEN<sup>1</sup>

<sup>1</sup>Energy Research Centre of the Netherlands, Petten, The Netherlands

<sup>2</sup>Energy and Sustainability Research Institute Groningen, University of Groningen, The Netherlands

### Article Outline

Glossary

Definition of the Subject and Its Importance  
Introduction

PEMFC Component Costs and Performance: Targets, Status, and Developments

Operating Conditions Leading to Performance Loss and Shortening the PEMFC Lifetime

Materials Degradation and the Relation to Performance Loss and Shortening the PEMFC Lifetime

Future Directions

Bibliography

### Glossary

**Bipolar plate** Forms the connection between MEAs in a fuel cell stack. The bipolar plate includes the gas flow channels and may also include cooling channels. Bipolar plates are also called flow plates.

**Degradation** The gradual loss of performance. Irreversible degradation is due to change of materials properties. Reversible degradation can be caused by non-optimal operating conditions. Quantitatively, the degradation can be expressed as a voltage decay rate.

**Durability** The capability of the fuel cell to operate in the operating window with limited loss of performance.

**Lifetime** The number of hours that a fuel cell can be operated in the operating window with a pre-defined performance loss relative to the initial performance.

**MEA** Membrane electrode assembly is the result of joining two electrodes and the electrolytic

membrane together. Usually, the gas diffusion media are considered to be part of the MEA.

**Operating window** The range of conditions in which the PEMFC can be stably operated. Within this operating window, performance can still depend on the conditions, but irreversible performance loss is limited. The operating window includes the modes of operation, such as start/stop events and load cycling.

**PEMFC** Proton exchange membrane fuel cell. The operating temperature is around 80°C. Cold start, below 0°C, is possible. For transport applications, the PEMFC is the fuel cell of choice.

**Robustness** The capability of the fuel cell to operate outside the operating window without a significant irreversible loss of performance.

### Definition of the Subject and Its Importance

The Proton Exchange Membrane Fuel Cell, PEMFC or PEFC, is in development for transport applications as well as for power generators ranging from a few Watts to tens of kilo Watts. Despite the fact that fuel cells have many advantages, such as a high conversion efficiency at partial load, clean exhaust gases, modular design and low noise production, their marketability will depend heavily on whether these fuel cells can compete with the incumbent technologies on performance, cost, and reliability in a specific application. For transport, the benchmark at present is the internal combustion engine, which has been mass-produced since 1908, and is characterized by high performance, high reliability, and relatively low cost.

### Performance

The internal combustion engine can be cold started within seconds and can be operated worldwide under all climate conditions. It has become quite standard to have 100 kW of engine power under the hood, enabling highway and uphill driving without any concession. A fuel cell vehicle will have to meet the same standards with regard to a fast cold start, availability of power without compromising the available space for the user under all foreseeable conditions. The power density of the complete fuel cell system must be 650 W L<sup>-1</sup> and

650 W kg<sup>-1</sup>, which translates to 2,000 W L<sup>-1</sup> and 2,000 W kg<sup>-1</sup> for the fuel cell stack [1].

### Lifetime and Reliability

Service intervals for passenger vehicles have dropped to once per 30,000 km or once per 2 years, and vehicles last 10–15 years without major revisions, running more than 250,000 km. Consumers are not used anymore to car breakdowns, especially not within the first 5 years of first ownership. Fuel cell vehicles will have to meet the expectations of today, rather than develop along the line of the internal combustion engine, that is, provide reliability comparable to that of the internal combustion engine of decades ago.

This means that at the time of mass-market introduction, their expected lifetime needs to be 5,000 h, with an end of life performance, which is at least 90% of the performance at the beginning of life. No external conditions, except for severe crashes, might lead to severe deterioration of the fuel cell performance. In practice, this means that the fuel cell system needs to operate between -40°C and +50°C ambient temperature, under all relative humidities.

### Costs

For the PEMFC to become a viable option for transport, the cost of buying and operating a fuel cell system, which comprises all parts at present not part of a conventional vehicle, must be comparable to that of an internal combustion engine. A higher investment cost of a fuel cell system can be offset by a lower fuel cost, determined by the cost of hydrogen and the fuel economy of the fuel cell vehicle. Although consumers are increasingly eager to take fuel costs or reduced CO<sub>2</sub> emissions into account when making their buying decision, there will be limits to the additional price they will want to pay. A cost-neutral switch from the internal combustion engine to a fuel cell system leads to an allowable cost of US\$30/kW for a complete PEMFC system and US\$15/kW for a PEMFC stack, which will be operated on hydrogen [1]. These costs are a reflection of the costs of the components and the assembly of the PEMFC stack and system, with the implicit assumption that it will last for the whole lifetime of the vehicle.

### The Role of Materials

The materials used in the PEMFC play a key role in the fuel cell systems performance, cost, and reliability. This article aims to present a comprehensive treatment of the performance, cost, and durability issues, especially but not exclusively, in light of their application in transport, as this provides so far the most challenging combination of these issues.

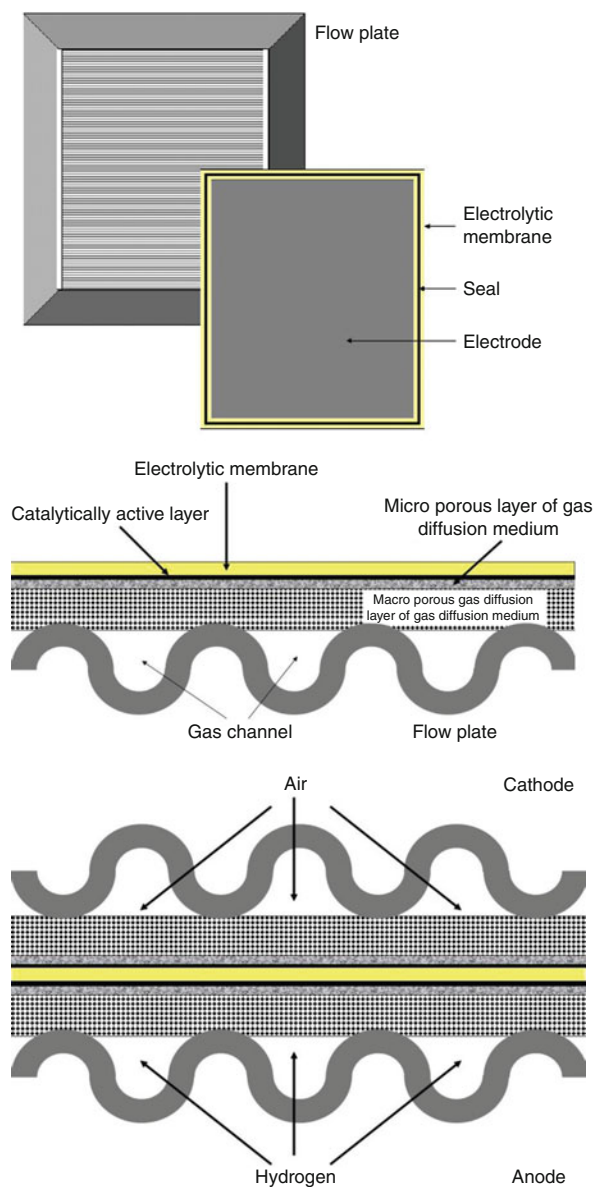
### Introduction

As of 2010, fuel cells have matured considerably: Hundreds of fuel cell-driven passenger vehicles have been demonstrated, with an impressive improvement of systems performance. Major car companies such as Honda, Toyota, Daimler, General Motors, Ford, and Hyundai have recently produced a new generation of passenger vehicles that can meet consumer expectations with respect to driving speed, acceleration, and driving range. Furthermore, they can be operated under severe conditions, such as extreme ambient temperature or demanding driving conditions. This accomplishment is based on a combination of advances on both fuel cell components as well as on systems level. On the one hand, the properties of materials determine the real power density of the fuel cell and the volumetric power density of the fuel cell stack. On the other hand, the systems layout, balance of plant and control strategy determine which conditions the fuel cell materials actually experience.

A schematic presentation of cell and stack components is given in Fig. 1.

The electrolyte of the PEMFC consists of a proton exchange membrane. Besides conducting ions from one electrode to the other, the electrolyte serves as gas separator and electronic insulator. At either side of the membrane, a catalytically active electrode is intimately attached to the membrane, to form a so-called membrane electrode assembly, the MEA. At the anode, hydrogen is oxidized to protons. At the cathode, oxygen is reduced to water. The theoretical voltage at open circuit of a hydrogen–oxygen fuel cell is 1.23 V at 298 K. Under load conditions, the cell voltage is between 0.5 and 1 V, producing a current density up to 1.5 A cm<sup>-2</sup>, depending on conditions.

The type of materials being used in PEM fuel cells has stayed the same since the early 1990s [2].



**PEM Fuel Cell Materials: Costs, Performance and Durability. Figure 1**

Cell and stack components. *Top*: top view of separator plate and MEA with seal; *middle*: cross-section view of one site of MEA on flow field; *bottom*: cross-section view of complete cell package

For the electrolyte, perfluorosulfonic acid/tetrafluoroethylene copolymer membranes (usually referred to as PFSA membranes) have been used for decades. Chemically, this polymer can be considered as

a Teflon or polytetrafluoroethylene (PTFE) backbone, with Teflon-like side chains bearing a  $\text{SO}_3\text{H}$  (sulfonic acid) group. Dissociation of this sulfonic acid group leads to mobile protons. Produced as thin, flexible sheets that become conductive when containing water, these membranes enable high volume manufacturing of complete cells by coating the electrodes on these sheets. Originally developed for chlor-alkali electrolyzers, the membranes provide a combination of properties that is unsurpassed: high proton conductivity and a high chemical stability. For optimal conductivity, the water content of the membrane needs to be such that the  $\text{H}_2\text{O}/\text{SO}_3\text{H}$  ratio  $\lambda$  is larger than 14 [3], which for state-of-the-art materials is only achieved at 100% RH. The latest developments on perfluorosulfonic acid/tetrafluoroethylene copolymer membranes that have been applied in fuel cells have been on even further improvement of their chemical stability and the development of ever thinner membranes down from 175 to 25  $\mu\text{m}$ , among others by using reinforcements for maintaining strength.

For the electrodes, platinum on carbon catalysts have long been used in both anode and cathode for hydrogen/air fuel cells. Commercial electrodes contain around 0.2–0.4  $\text{mg cm}^{-2}$  platinum, generating a power density of 0.5–0.7  $\text{W cm}^{-2}$  [2]. Using a total loading of 0.6  $\text{mg cm}^{-2}$  and a power output of 0.5  $\text{W cm}^{-2}$ , the platinum usage amounts to 1.2  $\text{g kWe}^{-1}$ . It has, however, been demonstrated that fuel cells with 0.4  $\text{g}_{\text{Pt}} \text{kWe}^{-1}$  are achievable when using clean hydrogen and air [4]. The long-term stability of such cells is, however, not known yet, and the use of reformat prescribes higher loadings of Pt–Ru at the anode, 0.2  $\text{mg}_{\text{PtRu}} \text{cm}^{-2}$  at minimum. The ultimate goal is to lower the platinum usage to less than 0.2  $\text{g}_{\text{Pt}} \text{kWe}^{-1}$ .

The use of noble metals is an important factor in the cost of the fuel cell. Whereas the cost of many components drops when the scale of manufacturing increases, this is not the case for the noble metal catalysts. The concern of a real shortage of platinum in case of large-scale use of fuel cells in vehicles has been proven not to be substantiated [5], but this is based on a significant reduction of its use to 15  $\text{g/vehicle}$ , corresponding to the 0.2  $\text{g}_{\text{Pt}} \text{kWe}^{-1}$  mentioned before. The key issue is to minimize the amount

of platinum per kW fuel cell power, while maintaining the power density of the present state-of-the-art. It makes no sense to substitute platinum by another metal when this leads to a reduction of the power density.

The catalysts used as a base for electrode manufacturing consist of high loadings of noble metal on carbon, of 40 wt% or even higher. These high loadings are used to render a thin electrode with high enough amount of active sites, typically 10  $\mu\text{m}$  thick. The platinum particle sizes are even at these high noble metal loadings in the range of 2–3 nm [6].

Even in electrodes with a high catalytic activity, the performance is heavily dependent on the electrode structure, as oxygen transport becomes crucial at practical current densities. Water removal plays a key role in this, as the diffusion constant of oxygen in water is a factor of 5,700 lower than that in air at 60°C. Gas Diffusion Media (GDM) play a decisive role in the water management of the fuel cells. Gas diffusion media, which consist of a macro porous gas diffusion layer (GDL) of 200–400  $\mu\text{m}$  covered by a micro porous layer (MPL) of 30–50  $\mu\text{m}$ , facilitate the transport of gas and electrons between the catalytically active layer and the flow plate. While the macroporous gas diffusion layer needs a certain thickness to distribute the gas in horizontal direction, the microporous layer facilitates the removal of liquid product water, preventing the so-called flooding of the electrode. Effective prevention of such flooding can extend the voltage current curve by around 0.5–1  $\text{A cm}^{-2}$ .

The component that has the highest impact on the weight and volume of the fuel cell stack is the flow plate or bipolar plate. Whereas the flow plates used to be made from high-density graphite, nowadays the material of choice is a moldable graphite/polymer composite material. Although the latter has a somewhat lower conductivity than pure graphite, it enables the use of plates with lower thickness due to its higher mechanical strength and its higher flexibility. This directly leads to reduction of stack weight and volume. A major advantage of polymer/graphite plates is the fact that they can be manufactured by means of injection molding [7]. An alternative to graphite and polymer/graphite material plates is the metal plate. The main advantage of

metal plates is the fact that very thin metal sheets can be used, and mass manufacturing techniques are available for forming flow patterns in these sheets [2].

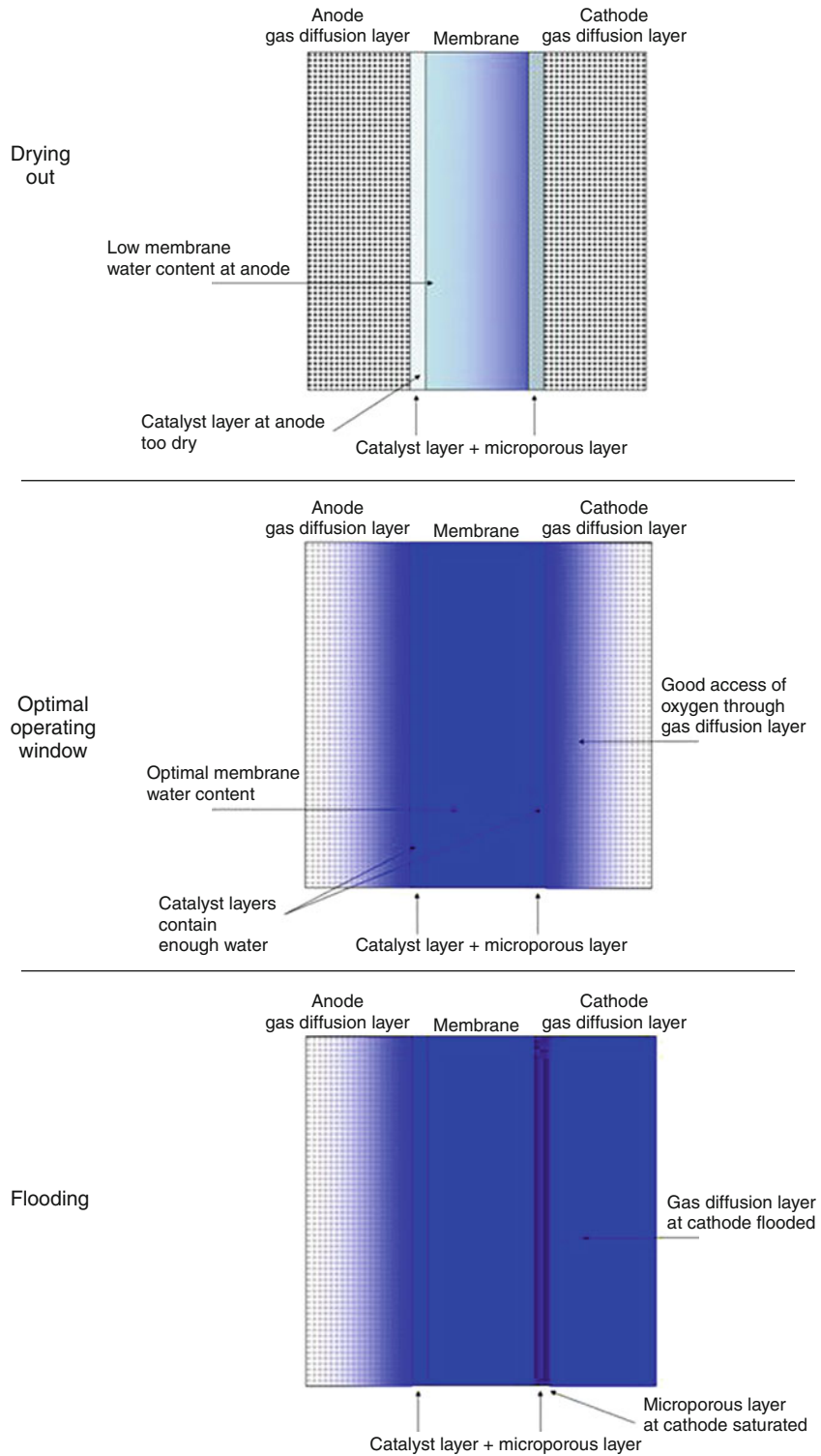
### Operating Window and Performance

As the area power density of a fuel cell is determining for the overall cost of the system, it is essential to optimize the operating window of the fuel cell stack. Of all conditions, the relative humidity has the largest influence on power density. This relative humidity is governed by the water content of the feed gases and their stoichiometry, the amount of product water, which is directly proportional to the current density, and the cell temperature.

The minimum requirement for water is set by the dependence of the proton conductivity of the PFSA membrane on its water content. In case this minimum requirement is not met, drying out of the membrane and the ionomer phase of the electrode leads to rapid decline of the power density. A maximum is clearly set by the transport of oxygen through the electrode, as liquid water is an effective barrier for the transport of gas to the reaction interface. In the catalytic layer of the cathode, these two factors determine an optimum: sufficient liquid water in the ionomer for proton conduction while simultaneously allowing gas phase transport of oxygen [8]. Figure 2 schematically draws the optimum situation as well as the situation of drying out and flooding.

The water accumulation at the cathode catalyst layer not only originates from product water but also from the electroosmotic drag, that is, protons migrating from anode to cathode also carry water with them. This electroosmotic drag is roughly proportional to the current density. Removal of this accumulated water can take place in two directions. First, water will be transported through the cathode gas diffusion medium to the air gas channel either as vapor or in the liquid form. Second, there will be back-transport of water through the membrane and the anode gas diffusion layer to the fuel gas channel. The respective rates at which these processes take place depend on the driving force for the water transport in either direction and the water permeability of the components [9].

The driving force is the difference in the chemical potential of water (determined by vapor pressure or



PEM Fuel Cell Materials: Costs, Performance and Durability. Figure 2

Water content of electrode parts and membrane in optimal operating window, as well as under too dry and too wet conditions



hydraulic pressure) in the catalyst layer and the gas channel. The first depends strongly on the current density, the second on the relative humidity of the gas channel, factors that can to some extent be controlled and matched.

The water permeability of the components is at least as important. Thin membranes allow for fast water transport to the anode, reducing flooding at the cathode. In thin membranes, the back-transport of water can often overcompensate the water transport that is associated with the electroosmotic drag, thus preventing drying out at the anode.

The state-of-the-art gas diffusion media are hydrophobized to such an extent that they allow transport of liquid water, an important mechanism at near-saturated conditions, as well as of water vapor and reactant gases. An important role is played by the micro porous layer (MPL). Because of the presence of small hydrophobic pores, a substantial liquid water capillary pressure can be built up, enabling a good gradient in the chemical potential of water to drier sections [10]. The optimization of gas diffusion media and the application of the MPL have led to significant improvement of the fuel cell performance at saturated conditions, showing their critical role.

The occurrence of flooding leads to an almost immediate drop in power output, which cannot be restored instantaneously. Proper design of the flow field, the cell characteristics, and knowledge of the operating window in which flooding as well as membrane drying can be avoided with the hardware used should lead to system control design avoiding these conditions. Sensors that measure the relative humidity continuously do exist [11] but seem to be too bulky and costly to be applied in a fuel cell system. In fact, the accurate measurement of the water content of gases is very complicated. Most systems therefore rely on the proper functioning of the humidifier, the cell temperature sensor, gas flow sensors, and the preservation of the water management properties of the gas diffusion media. If in the course of the cell lifetime, the water removal capability of the gas diffusion media declines, one could compensate for this by increasing the reactant stoichiometries at high current densities.

A more drastic prevention from flooding would be to work at drier conditions. In such a case, the conditions in the gas channel are such that the driving force

for water removal is enhanced. In addition to the strategy of developing membranes and catalysts that do not require a high humidity, the water permeability of the gas diffusion media should be reduced. Whereas mere reduction of gas permeability may also deteriorate the access of especially oxygen, improved water management may come from optimized MPLs. The capability to work at drier conditions would also allow increase of the upper limit of the window of operating temperature.

In the following sections, the fuel cell components are discussed in detail: options and needs for further cost reduction, for operation at more desired conditions, and durability issues.

## **PEMFC Component Costs and Performance: Targets, Status and Developments**

### **Cost and Performance Breakdown**

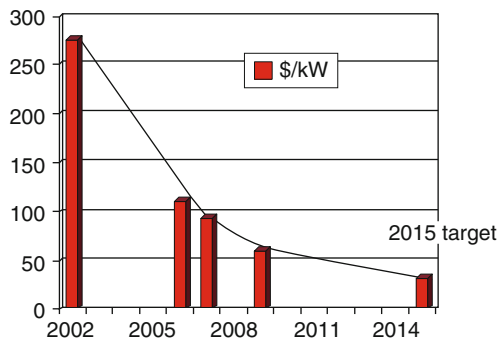
Directly or indirectly, all R&D efforts on PEMFC materials and components can be brought down to cost reduction. The progress on cost reduction over the last decade has been impressive. Based on high volume numbers, 500,000 systems per year, the cost of a PEMFC system using the materials available in 2009 would be \$61/kW coming from \$275 in 2002 [12]. These costs are a reflection of the costs of the components and the assembly of the PEMFC stack and system, with the implicit assumption that it holds for the rest of the lifetime of the vehicle. The durability requirement of 5,000 h for a passenger vehicle is treated separately.

For commercial applications, operation and maintenance costs are considered as important as the initial investment costs. Durability becomes in that case a matter of maintenance costs, both from the point of replacing individual components, as well as taking into account the time needed to replace these components. An individual seal in a fuel cell stack might cost only 1 dollar, but when the complete stack has to be disassembled to replace that individual seal, it is obviously worthwhile to apply more robust seals that minimize the need for replacement. Similarly, accepting higher fuel costs, caused by a lower cell voltage needed to generate the same fuel cell power after a certain voltage decay has accumulated over time, may be preferred to replacing MEAs or even a complete stack.

The most consistent monitoring of the progress on cost reduction is probably done by the US Department of Energy (DoE). Cost targets are set for 2010 and 2015, where the 2015 target is meant to meet commercial requirements, and the 2010 target is meant as intermediate milestone. An annual production volume of 500,000 vehicles per year is used to take into account the beneficial effect of mass manufacturing, which does not imply that a learning factor is incorporated in the cost figures. The cost per kW follows from the cost of components used per  $\text{m}^2$  cell area and the power density of such a cell under the specified conditions. The use of expensive components can thus lead to low costs per kW provided the power density of such a cell is high enough.

Figure 3 gives the progressive cost reduction of a PEMFC stack for the years for which enough details are present to make a deeper assessment of the cost breakdown.

It must be noted that the DoE analysis has some tentative aspects. New materials are taken into account in the cost figures without the guarantee that these will actually survive the operating conditions over the full lifetime of the fuel cell. On a cell level, these materials might have survived relevant accelerated tests, but that does not mean that in combination with stack and system hardware, and used under real-life conditions, these materials will qualify. It is at present the only publicly known assessment of fuel cell costs that is



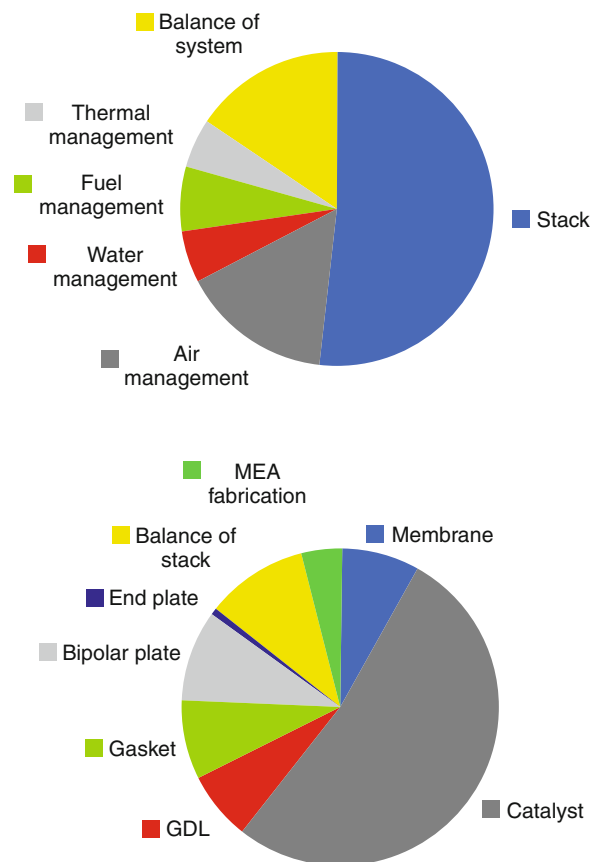
PEM Fuel Cell Materials: Costs, Performance and Durability. Figure 3

Cost progress for PEMFC stack, \$ per kW for an 80 kW stack, when manufactured at 500,000 pieces per year (Data from [12])

regularly updated, and that takes into account all fuel cell stack and system components on an equal basis.

According to the most updated report [13], the cost of an 80 kW PEMFC system, when produced in a quantity of 500,000 systems per year, would amount to \$60–\$80/ $\text{kW}_{\text{net}}$ . The contribution of the stack and systems components is illustrated in Fig. 4.

From Fig. 4, it becomes clear that the stack is the most expensive single component of the fuel cell system, and that on stack level, the catalyst is the most expensive single component. It is important to notice that all costs are expressed per kW system output. It can



PEM Fuel Cell Materials: Costs, Performance and Durability. Figure 4

Contribution of components to total stack costs, taking the cost estimate from Tiax [13]. Top figure: cost breakdown of an 80 kW system; bottom figure: cost breakdown of an 86 kW PEMFC stack. Note that an additional 6 kW is necessary to power balance of system components

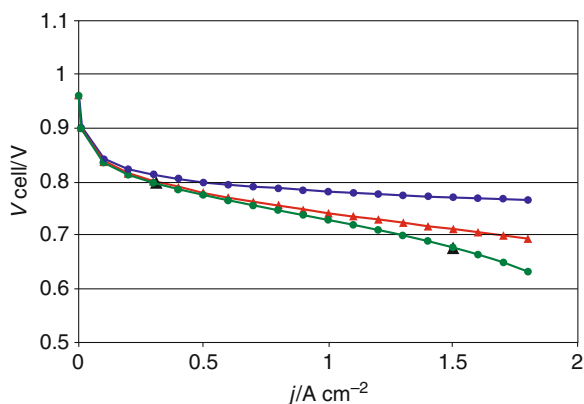
therefore be important to improve a component that has a big impact on fuel cell performance without being expensive itself. When a better performing membrane or gas diffusion layer leads to a doubling of the power density while the catalysts cost per square meter stays the same, its cost will halve per kW due to this membrane or GDL improvement. Vice versa: It is not helpful to decrease the loading of platinum per square meter, when this leads to a lower power density: All components' costs per kW will increase, possibly offsetting the cost gain of the lower platinum cost.

In PEMFC, the cost targets strongly determine the performance targets, which includes MEA performance as well as the allowable operating window. As a result, targets are more demanding for automotive than for stationary applications.

For automotive conditions, the DoE has set for 2015 a target rated cell power density of  $1 \text{ W cm}^{-2}$ , that is,  $1.5 \text{ A cm}^{-2}$  at  $0.68 \text{ V}$  corresponding to 55% LHV electrical efficiency. The requested efficiency at 25% power is 65%, which implies  $0.8 \text{ V}$  at  $0.25 \text{ W cm}^{-2}$  or  $0.31 \text{ A cm}^{-2}$ . Furthermore, the cost requirements demand that the total precious metal loading should not exceed  $0.15 \text{ g kW}^{-1}$ , which results in  $0.2 \text{ mg cm}^{-2}$  [1]. In automotive conditions, the peak power will be achieved at the higher end of the temperature window, in extreme situations near  $120^\circ\text{C}$ , which will mean that the relative humidity will be low, in the order of 25% maximum.

In stationary systems (including APU), the power density may be lower, with  $0.6\text{--}0.7 \text{ W cm}^{-2}$  at  $0.7 \text{ V}$  a typical target but with higher precious metal loadings in the order of  $0.5 \text{ mg cm}^{-2}$ . For reformat fed systems, there is the additional requirement of CO tolerance up to 50 ppm. The temperature and relative humidity can be kept closer to what are called the ideal PEMFC conditions, although increase of temperature up to  $120^\circ\text{C}$ , without increasing the dew point of the gases, would contribute to efficiency of the utilization of heat in  $\mu\text{CHP}$  systems. As automotive performance targets seem to encompass stationary targets with the exception of reformat tolerance, in the following, the emphasis will be on automotive targets.

Figure 5 shows a deconvolution of the total cell voltage compared to the voltage corresponding to the Lower Heating Value. Three different types of losses are



**PEM Fuel Cell Materials: Costs, Performance and Durability. Figure 5**

Deconvolution of the total cell voltage according to DoE targets. The blue line includes activation losses only, the red line includes ohmic losses as well, and the green line gives the overall performance in agreement with the DoE targets ( $\blacktriangle$ )

usually identified, ohmic loss, activation loss and transport losses. The curves shown correspond to the target situation, that is, the overall ohmic resistance is  $0.04 \Omega \text{ cm}^2$ , the total performance matches the efficiency requirements at rated power and 25% rated power, and the activation losses are limited to what may be expected if the catalysts satisfy the targets of the DoE, that is, a mass activity for the oxygen reduction reaction (ORR) of  $0.44 \text{ A mg}^{-1}$  ( $900 \text{ mV}$ ,  $\text{H}_2/\text{O}_2$ ,  $80^\circ\text{C}$ , 100% RH, 150 kPa), and  $0.15 \text{ mg}_{\text{Pt}} \text{ cm}^{-2}$  cathode loading [1]. At  $0.05 \text{ mg}_{\text{Pt}} \text{ cm}^{-2}$  anode loading, anode losses are assumed to be negligible. The difference between the targeted curve and the curve representing activation and ohmic losses reflects the maximum acceptable transport losses.

The most significant contributions to the ohmic losses are due to the membrane, the bipolar/cooling plates and the electrodes (GDM + catalyst layer), including contact resistance between components. In a  $\text{H}_2/\text{air}$ -fed fuel cell, the activation losses are mainly at the cathode. The catalyst layer, the microporous layer (MPL), and the macroporous gas diffusion layer (GDL) of the gas diffusion medium (GDM) as well as the gas channel design (i.e., the bipolar plate) all contribute to transport losses of reactants.

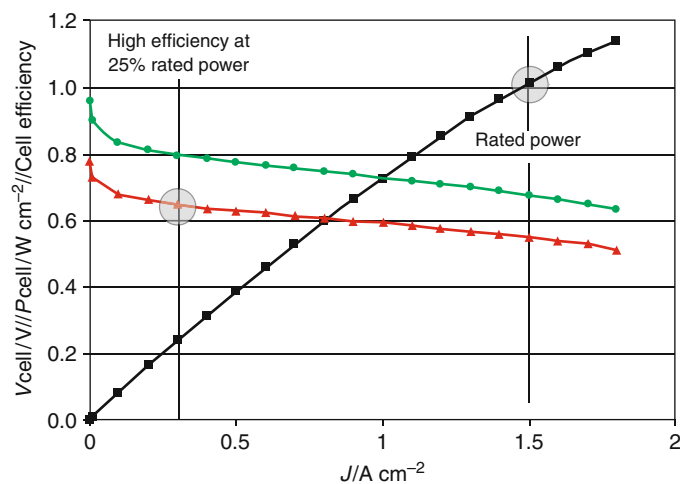
The PEM fuel cell contains a number of components of which the performance varies with operating conditions. This relation can be either instantaneous or become apparent after longer exposure to such conditions. The components that are most sensitive for operating conditions are the proton exchange membrane and the electrodes for hydrogen oxidation and oxygen reduction.

The PEMFC based on perfluorinated sulfonic acid membranes can be operated from sub-freezing conditions to around 80°C, the upper limit depending on the operating pressure and relative humidity of the inlet gases. Although the PEMFC can be started subzero, systems will generally be controlled in such a way that the cell temperature is allowed to increase to the desired set-point between 60°C and 80°C, where power output is at its maximum while improving water management. At low pressures and low relative humidity, this set-point will be close to 60°C, while at high pressures and high relative humidity, the set-point will be close to 80°C.

It has been amply demonstrated that the membrane performance is at its maximum when fully wetted by water. The proton conductivity is in this case at its maximum. Under stationary conditions, using fully humidified gases and moderate temperatures not

exceeding 70°C, the stability of the membrane was proven sufficient to sustain stable performance for 26,000 [14] – 36,000 [15].

For the fuel cell electrodes, the picture is more complex, partially because of the various electrode components and functions merged together. For proton conductivity, an ionomer similar to that used for the proton-conducting membrane is dispersed throughout the electrode. As for the membrane, it needs to be wet to provide sufficient proton conductivity. At the same time, an upper limit for wetting exists for the catalyst layer to prevent flooding. For the electrocatalyst, the electrode potential is the dominant factor determining activity and stability, provided that hydrogen and oxygen are sufficiently available at the reaction interface. For hydrogen oxidation, the activation losses are moderate, so that in practice the anode is between 10 and 50 mV versus RHE. For oxygen reduction, however, the activation losses are substantial; the cathode potential is, depending on the current density, between 500 and 800 mV versus RHE. There has always been a trade-off between efficiency and power output, which can be translated into a trade-off between fuel efficiency and fuel cell investment cost. The relation between cell current density and cell voltage, efficiency, and power density is illustrated in Fig. 6.



PEM Fuel Cell Materials: Costs, Performance and Durability. Figure 6

Cell voltage (*green*), power density (*black*), and efficiency (*red*) versus cell current density. The points of rated power and 25% rated power are indicated. Follow vertical lines to find the corresponding cell characteristics at these points

The fuel cell electrical efficiency is directly proportional to the fuel cell voltage  $\Delta E_{\text{cell}}$ :

$$\text{Eff}_{\text{FC, LHV}} = \frac{\Delta E_{\text{Cell}}}{1.23 \text{ V}}, \quad \text{or} \quad \text{Eff}_{\text{FC, HHV}} = \frac{\Delta E_{\text{Cell}}}{1.48 \text{ V}}$$

for the lower heating efficiency (LHV) and the higher heating efficiency (HHV), respectively, at room temperature. In Fig. 6, the green line represents the relation between cell voltage and current density, which is the same as in Fig. 5; the red line represents the LHV cell efficiency versus current density, according to the equation above, while the black line represents the cell power density versus current density. For the fuel cell and conditions of Fig. 6, the point of rated power density, in this case  $1 \text{ W cm}^{-2}$ , lies at 0.68 V, at which the cell LHV efficiency amounts to 55%. At a cell LHV efficiency of 65%, the power density equals  $0.25 \text{ W cm}^{-2}$  at a cell voltage of 0.8 V. In practice, a system containing a cell with characteristics shown in Fig. 6 will be operated near the high efficiency point for most of its time, while at occasional times, when the rated power is needed, the cell is operated at lower cell voltage. Opposed to internal combustion engines, the operation at low power demands, for example, when idling, leads to higher efficiencies of the fuel cell. At systems level, this is partly offset by the energy consumption of balance of system components.

When developing new materials and components for reducing costs, their durability should be the starting point instead of the sanity check afterward. The durability standard is set by the components used in today's technology, and should only become better. Cheaper alternatives that would lead to lower durability are not welcome, as they would lead to the previously mentioned early replacement of complete stacks or disassembling of the stack to replace individual components.

Moreover, cheaper cell materials leading to lower power density will lead in most cases to higher stack costs at the end. Except for the catalyst, no single component constitutes more than 10% of overall stack costs. Using the overall cost numbers of the Tiax analysis [13], total cost per cell will be \$4,80 including bipolar plates. The power of such a single cell is 196 W, that is, costs are \$24.50 per kW. The membrane, generally regarded as an expensive component, will cost \$0.43 per single cell. Suppose a cheap alternative

membrane, costing only \$0.10 per cell, would lead to a performance of 180 W per cell. In this case, a very substantial cost reduction of more than 75% for a particular component leads to a power output loss of less than 10%. Per kW, the cell with the low cost membrane is more expensive, \$24.8 per kW than the cell with the expensive membrane. This simple example shows that improvement of the cell power density rather than mere component cost reduction should be driving materials R&D.

The complexity of the fuel cell system is for a considerable part caused by the large influence of the relative humidity in the cell on the performance of the MEA. The heat removal capacity of a fuel cell system operating at 70–80°C is limited, since it is directly proportional to the temperature difference between the cooling liquid temperature and the ambient temperature. Moreover, nearly all the heat produced in the fuel cell stack has to be rejected over the coolant/radiator/ambient air heat exchanger, while in the internal combustion engine, much heat is rejected through the tailpipe. Under certain conditions, that is, when delivering full power at a limited speed at high ambient temperature, this can easily lead to stack temperatures up to 120°C. Volkswagen has a test drive in Death Valley to test their cars on their capability to deal with this [16]. Although the occurrence of such events can be minimized by increasing, for example, the radiator size of the system, from a point of view of system robustness, it is preferred that the stack can operate at such conditions.

However, maintaining a constant high relative humidity at temperatures rising above 70–80°C is not feasible due to the large amount of water that would need to be circulated as well as the negative effect it would have on the oxygen partial pressure. Therefore, a large R&D effort is currently spent on developing MEAs that do not show a drop in power output at increasing temperatures with concomitant reduction in relative humidity. Although such widening of tolerable relative humidity range could lead to some reduction in cost for thermal management and water management, see Fig. 4, the prime advantage is on increasing robustness of the system.

In the following sections, the individual cell and stack components are discussed in view of their performance and cost.

## Costs and Performance Developments of Components

**Membranes** The targeted values for the membrane resistance are  $0.02 \Omega \text{ cm}^2$  [1]. The accepted commercial standard for membranes (Nafion, Flemion, Aciplex,) consists of a PTFE backbone with a perfluorovinylether side chain that ends with a sulfonic acid group. Variations in this side chain distinguish the different trade names. The commercial standard membranes have equivalent weights (EW) in the order of  $1,100 \text{ g eq}^{-1}$  (weight per mol sulfonic acid groups). At ideal conditions (i.e.,  $80^\circ\text{C}$ , liquid water present), a conductivity of  $0.1 \text{ S cm}^{-1}$  is a standard value for commercial PFSA membranes. The resistance target therefore corresponds to a maximum thickness of  $20 \mu\text{m}$ . Whereas this was a few years ago still at conflict with requirements on gas tightness and durability, the implementation of reinforced membranes has made this feasible as was shown by Gore, Dupont, and 3M [17–19]. Such a reinforcement is made of PTFE or a ultrahigh molecular weight polyethylene. The latter materials, however, are less suitable for operation near  $120^\circ\text{C}$ . These high EW PFSA membranes of  $25\text{--}30\text{-}\mu\text{m}$  thick, either reinforced (DTI) or not (TIAX), form the basis of the membrane cost in the DoE analysis. These costs amount to  $\$2.4$  (TiAx) or  $\$3.3$  (DTI) per kW [13].

The conductivity of high EW membranes, however, drops dramatically with relative humidity to less than  $5 \text{ mS cm}^{-1}$  at 25% RH, that is, by more than an order of magnitude [20]. Although conductivity values increase slightly with temperature, this implies that at  $120^\circ\text{C}$ , 25% RH these membranes do not meet the target. This fact has motivated considerable research into membranes that can operate at drier conditions, with already quite promising results.

Several routes have been and are to time still being explored, which include (1) low EW PFSA membranes, (2) hybrid inorganic/organic membranes where inorganic additives should take care of water retention and/or proton conductivity, and (3) membranes that enable intrinsically dry proton conduction, such as systems based on imidazole or phosphonic acid rather than water as a proton carrier or relying on phosphoric acid as the proton-conducting medium.

Low EW membranes have higher proton conductivity values at a given humidity. The equilibrium

number of water molecules per acid group ( $\lambda$ ) at a certain humidity value is not dependent on the EW, which results in an increased water uptake per volume or weight unit. In addition, the mobility of the protons tends to increase with decreasing EW, supposedly due to morphological changes associated with the higher water fraction [21]. The combination of higher concentration of protons and enhanced mobility results in conductivity values that at the same  $\lambda$  are higher for low EW polymers. At low EW, however, the degree of crystallinity of the membrane is reduced, which results in water-soluble membranes. Approaches to reduce the water solubility have included modification of the polymer, cross linking, and blending.

A modification of the polymer that has been adopted by various groups is to have shorter side chains as compared to Nafion. Short side chains increase the crystallinity of the PFSA, thus reducing the solubility. Solvay Solexis has developed Aquivion, a membrane based on Hyflon, which is a copolymer of Teflon and sulfonyl fluoride vinyl ether with low EW (790–870) and good crystallinity, with proton conductivity values in the order of  $30 \text{ mS cm}^{-1}$  at  $120^\circ\text{C}$ , 30% RH [22]. A similar approach is followed by 3M, who have shown 580 EW membranes approaching  $100 \text{ mS cm}^{-1}$  at  $120^\circ\text{C}$  and RH 50% [23]. Gore recently reported values  $>50 \text{ mS cm}^{-1}$  at 30% RH and  $>100 \text{ mS cm}^{-1}$  50% RH with a new, undisclosed ionomers [24]. DuPont recently presented results on MEAs with new ionomer that showed a much reduced dependence on the RH as compared to Nafion-based membranes [17].

Cross-linking can be achieved through the backbone, but also through the acidic groups. Dongyue Shenzhou New Materials uses sulfonimide links to this end [25]. Values of  $200 \text{ mS cm}^{-1}$  at  $80^\circ\text{C}$  and 95% RH to  $12 \text{ mS cm}^{-1}$  at  $120^\circ\text{C}$  and 25% RH have been reported with this material. Also 3M is considering introducing (aromatic) imide groups to the sulfonic end group either to attach cross-linkable groups or more acid groups per side chain [23]. So far, approaches involving blending of soluble and insoluble material do not seem to have been successful in stopping the dissolution of the soluble material. Reinforcements may also reduce the solubility of low EW materials.

The durability aspects, however, have not been fully established. These will be discussed in section

### “Materials Degradation and the Relation to Performance Loss and Shortening the PEMFC Lifetime.”

Also the addition of inorganic materials to enhance water retention and/or proton conductivity has had some success, for example, by adding zirconium phosphate. Such additions have also been successful in increasing the mechanical strength of materials [26, 27].

Hydrocarbon membranes, such as sulfonated PEEK or sulfon imides, usually show lower conductivity values and increased swelling, or water solubility. The proton conductive paths in these materials are not as effective as in PFSA membranes. Not unrelated, however, is the observation of much reduced gas cross-over rates in such membranes [25, 28]. While this may be unfavorable for the application of such materials as proton-conducting phase in the electrode, it has a positive impact on the durability of the membrane as will be discussed in section “Materials Degradation and the Relation to Performance Loss and Shortening the PEMFC Lifetime.” This has recently given a new incentive to research in this area [29]. Sulfonated PEEK membranes are commercialized by Fumatech GmbH in Germany [30]. Low equivalent weight membranes (EW 675–850) with a conductivity ranging from 0.04 S cm<sup>-1</sup> at 40°C to more than 0.08 S cm<sup>-1</sup> at 80°C are offered.

The success of widening the operating temperature for water-based proton conduction system has so far not been matched in the field of intrinsically dry proton conductors. Although interesting developments were reported in the field of polymers having tethered imidazole groups [31–34], these and other approaches based on phosphonated polymers, either in their pure form or as acid–base copolymers or blends, show conductivity values, which are at least an order of magnitude too low [34–36].

The proton conduction based on the phosphoric acid is the basis of HT-PEMFC Celanese technology [37], mostly referred to as phosphoric acid-doped PBI (polybenzimidazole) This membrane enables operation at temperatures as high as 180°C, without the need for external humidification. Heat dissipation at this temperature is much easier than at the 70–80°C operating temperature of fuel cell systems using standard PFSA membranes. The CO tolerance at 180°C is such that even 1% CO leads to a minor loss of power

density compared to that using the same membrane on pure hydrogen. The downside of this membrane is its low conductivity below 100°C, making a cold start impossible, as well as the lower power density at its optimal temperature.

At 23°C, the power density of a phosphoric acid-doped PBI-based MEA is quoted to amount to 8 mW cm<sup>-2</sup> [38], while the best performing MEA at 160°C shows a power density of 0.28 W cm<sup>-2</sup> at 0.7 V [39], less than half of the PFSA-based MEAs used for the DoE cost analysis. The only car manufacturer pursuing the use of phosphoric acid-doped PBI membranes has been Volkswagen. For stationary applications, it has been primarily PlugPower that has developed fuel cell systems based on phosphoric acid-doped PBI membranes.

Unassisted start-up at freezing conditions, another automotive requirement, does not seem to be compatible with systems not based on water. First, as already mentioned, their proton conductivity is too low at such temperatures. Secondly, they cannot play the role of water storage medium during start-up. By pre-drying PFSA membranes, the water content can be reduced so much that only non-freezable water remains. This results in a remaining conductivity in the order of 10 mS cm<sup>-1</sup> that is sufficient for a start-up procedure. The low initial performance induces heat generation. In a previously dried system, product water can then be stored, keeping the gas channels in the active layers free from ice [40, 41]. This start-up procedure is not feasible for membranes unable to adsorb liquid water, such as PA-doped systems where water adsorption would lead to washing out of the acid.

In a cost comparison by Gebert et al. [42], it is concluded that sulfonated Polyether ether ketone (S-PEEK) and phosphoric acid-doped polybenzimidazole (H<sub>3</sub>PO<sub>4</sub>-PBI) could be a factor of five cheaper than Nafion. As discussed earlier, this cost advantage is only helpful when not offset by lower MEA power density.

**Electrodes and Gas Diffusion Media** As noted in the expert review of the two parallel cost estimates for the DoE program [13], the cost reduction realized in the past 3 years is almost entirely caused by the reduction in platinum loading to 0.25 mg cm<sup>-2</sup> at an areal power density of 0.715 mW cm<sup>-2</sup>. It takes into account

a prescribed platinum cost of \$1,100 per troy ounce (= 31.1 g), to avoid a high volatility of fuel cell cost caused primarily by the volatility of the platinum price. At the same time, the high impact of the platinum price on the fuel cell cost is a serious problem: As the loading target is distilled from the platinum price and the allowable cost per kW, a successful reduction of the platinum loading can easily be offset by an increase in the platinum price. A further reduction of the platinum loading could alleviate this dependence.

The catalytic layer used for the DoE cost review [13] is based on a ternary PtCoMn alloy, either supported by carbon or by organic whiskers, as developed by 3M, with a platinum loading of  $0.25 \text{ mg cm}^{-2}$  cell area. The cost of the support and the ionomer is insignificant compared to the precious metal cost of the catalytic layer. Most R&D is devoted to decreasing the use of precious metal, primarily platinum, while at the same time preserving power density and durability.

The PtCoMn alloy as developed by 3M that was adopted for the cost review is so far the only catalyst for which the activity data set by the DoE have been met in MEA tests under relevant conditions, but long-term field data are absent for stacks or systems based on the PtCoMn alloy. The most applied cathode catalyst to date is Pt supported on carbon, while at the anode, either Pt or Pt alloyed with Ru and or Mo is used. At present, these catalysts do not meet yet the target activity.

Further electrode performance improvements, and thus reductions in use of platinum per kW, can either come from a further reduction of ohmic losses, activation losses, or from transport losses.

*Ohmic losses:* The ohmic losses associated with the electrodes are related to electron transport through the gas diffusion media and the catalyst layers. The proton transport in the catalyst layer is associated with transport losses, as its contribution is not ohmic but depends on current density [43]. In state-of-the-art components, carbon is the electron conductor.

The in-plane as well as the through-plane resistance of the GDM contributes to the ohmic resistance due to their function of evening out the current distribution. As most current GDM's are based upon non-isotropic materials such as paper or woven materials, these differ often by an order of magnitude. The through plane resistance of the material depends strongly upon the

compression, a factor that has to be taken into account for application in a fuel cell. In general, the overall ohmic loss through the GDM is quite small, in the order of  $2 \text{ mV}$  at  $1 \text{ A cm}^{-2}$  [44].

The presence of a so-called microporous layer at the interface with the catalyst layer has a positive effect on the contact resistance. For carbon black-based catalyst layers, it is usually assumed that electron transport losses are negligible. This may not be the case when less well-conducting oxide or carbide supports are considered.

*Activation losses:* In an  $\text{H}_2/\text{air}$ -fed fuel cell, the activation losses are mainly at the cathode. The hydrogen oxidation rate is very fast on the standard Pt catalyst, indeed such that it is hard to measure it accurately. Estimates of the exchange current density are in the order of  $0.24\text{--}0.60 \text{ A cm}_{\text{Pt}}^{-2}$  [45]. There is substantial evidence that the anode Pt loading can be as low as  $0.05 \text{ mg cm}^{-2}$  with losses in the order of  $\text{mV}$  only [4]. Durability issues as well as the presence of CO may require a higher loading; this will be further discussed in the next chapters.

The research is therefore focused at the cathode. The state-of-the-art catalyst Pt/C shows only a low specific activity in the order of  $0.2 \text{ mA cm}^{-2} \text{ Pt}$  at  $900 \text{ mV}$  (IR-free, at 1 bar  $80^\circ\text{C}$ ) This is compensated by the large Electrochemically Active Surface Area values (ECSA) obtained with these catalysts, which can be in the order of  $60\text{--}90 \text{ m}^2 \text{ g}^{-1}$ . The corresponding mass activities, which are the product of the ECSA and the specific activity, are between  $0.12$  and  $0.18 \text{ A mg}^{-1}$ , about a factor 3 lower than the target, resulting in a required Pt loading in the order of  $0.4\text{--}0.6 \text{ mg cm}^{-2} \text{ Pt}$  [46].

A further increase of the ECSA does not seem useful; this would imply particles smaller than  $2 \text{ nm}$ , which are not very active. Specific activity decreases with particle size or specific surface area due to more low coordinated Pt atoms and longer Pt–Pt bond distances, which have a pronounced effect on the electronic structure. Ensuing strong OH adsorption blocks the sites for oxygen adsorption. Whereas  $0.2 \text{ mA cm}^{-2}$  seems a sort of maximum for Pt nanoparticles, the maximum specific activity for pure Pt with well-defined crystal faces is much higher, in the order of  $2.2$  and  $1.9 \text{ mA cm}^{-2}$  for Pt (110) and (111), respectively, measured in  $\text{HClO}_4$ , an electrolyte that due to its non-adsorbing anions



mimics PFSA best [47]. For Pt black and Pt polycrystalline disk, higher values for the specific activity are also found [46]. 3M showed with their so-called Pt Nano-Structured Thin Films (NSTF) a specific activity of up to  $1.7 \text{ mA cm}^{-2}$  Pt at 900 mV (IR-free, at 1 bar  $80^\circ\text{C}$ ) with ECSA values in the order of  $10\text{--}15 \text{ m}^2 \text{ g}^{-1}$  [48, 49]. The NSTF is a continuous layer of polycrystalline Pt deposited on non-conductive, inorganic whiskers [50], with a surface mostly showing Pt (111) facets [51].

Several ways to improve the mass activity are currently explored. In the concept of nanoparticles on carbon black, alloying Pt with other metals is intensively studied. It is well known that the specific activity of Pt can be improved by alloying. The effect is ascribed mostly to changes in the electronic structure, which bring the system nearer to the optimum where both O–O bond dissociation and OH formation take place [52, 53]. Very high specific activity values were recently reported for certain  $\text{Pt}_3\text{M}$  alloys ( $\text{M} = \text{Ni, Co, Fe}$ ), where the activity of  $\text{Pt}_3\text{Ni}$  (111) was shown to be ten times that of Pt(111), that is,  $19 \text{ mA cm}_{\text{Pt}}^{-2}$  [53, 54]. These findings have motivated studies into supported Pt binary and ternary nano-sized catalysts as well as extended Pt alloy catalysts.

Nano-sized binary alloys of Pt with Cr, Co, Mn, Ti, Co have been widely studied indicating a 3–4 fold increase in specific activity, but also identifying serious stability problems [6, 46, 55, 56]. More recently, ternary systems including PtIrCo, PtIrCr, and PtMnCo are considered with the aim to improve durability. Mass activity values of some PtIrM catalysts come close or exceed the  $0.44 \text{ A mg}^{-1}$  target value [57]. Increased activity can also be achieved by Pt skin nanoparticles. Here the core of a particle is made of cheap transition metal with a skin of Pt. By carefully selecting the base material, the skin Pt atoms may be more active than the Pt on the outside of a pure Pt nanoparticle [58]. High mass activities have been reported for particles consisting of a Pt or PtIr monolayer on  $\text{Pd}_3\text{Co}$  cores (order  $0.7 \text{ A mg}^{-1}\text{Pt}$ ) [59].

A relatively new class of catalysts are pre-leached alloys, e.g., Cu–Pt. It was found that after leaching of Cu from the outer layers of  $\text{Pt}_{0.25}\text{Cu}_{0.75}$ , the remaining  $\text{Pt}_{1-x}\text{Cu}_x$  showed a Pt-rich surface with high activity, again most likely due to favorable Pt–Pt distances and related electronic effects [60, 61]. Also Pt-free precious

metal alloys, such as Pd with a transition metal, are currently being investigated [62].

Changing the support material may also have an effect on the mass activity. In many cases, it was found that on low surface area carbon supports, which are expected to be more stable, the Pt cannot be so well dispersed resulting in lower specific surface area and mass activity. However, it may be envisaged that either through inducing a certain particle morphology or by an electronic interaction, the specific activity may increase in such a way that this outweighs the loss of surface area. Examples of this have recently been reported for Pt on nitrogen-modified carbon [63]. Other such effects have been reported for WC and VC supports. Other alternative supports include conducting oxides, which can also have an enhanced electronic effect [63, 64].

Non-supported catalysts seem a way not only to avoid durability issues related to the support (such as carbon corrosion), but also introduce extended surfaces and therefore the possibility of higher activity. The NSTF is an example, but also improved Pt black, Pt nano-wires or tubes, Pt-coated carbon nanotubes, or mesoporous Pt structures. Although such structures have been made, they have, with the exception of Pt black and NSTF, not yet been extensively tested in fuel cells. The high mass activity of NSTF Pt was already mentioned. By using extended alloys, that is, NSTF of PtCoMn, the DoE target has been exceeded for PtCoMn ( $>0.44 \text{ A mg}^{-1}$ ) [49].

The last class to mention are non-metal catalysts. The use of catalysts not using metallic catalysts at all has intrigued many researchers for long time. Inspiration has come from nature, where hemoglobin structures in mammals bind oxygen, while porphyrin structures do so in plants. The most promising class of non-metal catalysts are the composite catalysts derived from heteroatomic organic precursors (e.g., polyaniline, polypyrrole, cyanamide, etc.), transition metals (Co or Fe), and carbon [65, 66]. The mechanism has not been completely clarified, but it is becoming increasingly clear that four M–N bonds are required. An essential step in the preparation is heat treatment at  $600\text{--}1,100^\circ\text{C}$  and subsequent activation steps. A target here is a volume activity of  $130 \text{ A cm}^{-3}$  at  $0.80 \text{ V-IR-free}$ . This target is not so much derived from cost considerations but from the requirements on the

dimension of the electrode [46]. Still, here stability problems are reported as well as intrinsic mass transport losses [66, 67].

The effect of the humidity on activity has not been well explored due to a lack of suitable systems. It has been reported that activity drops when  $RH < 50\%$ , due to reduced proton activity [68]. Key to a high mass activity is the utilization of the catalyst, which is determined by accessibility for reactants. Under well-humidified conditions, supports with small pores show good utilization due to the presence of liquid water, which transport protons. This becomes critical at low humidity although smaller pores dry out less easily than larger.

*Transport losses:* The catalyst layer, the MPL, the GDL and the gas channel design (i.e., the bipolar plate) all contribute to transport losses of reactants. In the catalyst layers, both proton transport as well as transport of reactants takes place. At the anode side, transport losses are usually negligible. Hydrogen transport is fast, and the reaction seems to take place close to the membrane interface. At the cathode, the problems can be considerable due to the lower partial pressure of oxygen in air and the water accumulation. This implies the reaction does not just take place at the interface, and therefore, proton transport also plays an important role. Transport losses increase with current density. The oxygen gain, which usually gives a fair indication of mass transport losses, can be as high as 150 mV at  $1.5 \text{ A cm}^{-2}$  [69].

The proton resistance of the catalyst layer can be reduced by adding more ionomer phase or low EW ionomer to the system, but this is outweighed by reduced oxygen transport. As gas phase diffusion is several orders of magnitude faster than diffusion through liquid water, it is essential to create a system that does not completely fill with liquid water, that is, contains fairly hydrophobic materials and not too small pore sizes. The state-of-the-art carbon blacks do not seem to meet this criterion. CNT or other more graphitic structures seem better suited. Also alternative supports should be selected with their potential for improved mass transport. Oxides may be less suitable in that respect but this certainly requires further investigations. At low RH, the requirements may be different, to some extent the structure needs to retain water for proton conductivity, as well as for proton activity.

It has been observed that dry ionomers of the PFSA type are not very well permeable for gases. Reduced catalytic activity and more specifically increased transport losses are the reason why present state-of-the-art MEAs still have a poor performance at low RH and high T, in spite of acceptable ohmic resistance of the membrane. This requires improved ionomer to be used also in the electrode [70].

In general, thin catalyst layers will have the lower proton resistance. However, they may also fill up more easily with water as is, for example, the case with the very thin NSTF electrodes by 3M, which only seem to function well at non-saturated conditions. Also for start-up under freezing conditions, a thicker electrode, or at least a higher pore volume, seems to be an advantage as complete filling with ice is even more detrimental.

A certain amount of mass transport problems must be ascribed to the GDL and MPL. The presence of an MPL usually prevents the GDL from becoming saturated with water; in most state-of-the-art GDL's, the diffusion through the microporous layer is not limiting, although this may change upon aging. The MPL, although relatively dense, improves the transport by optimizing the water management [10]. When the fuel cell is to be operated at relatively dry conditions, water retention is more important than water removal. Also in this case, the GDM plays a critical role [71].

The cost of Gas Diffusion Media (MPL + GDL) [13] is quoted to be around \$10–\$15 per  $\text{m}^2$ , and is likely to be primarily determined by processing costs. Both carbon/graphite paper as woven structures are being used as substrate, which is made hydrophobic by, for example, a PTFE coating. The microporous layer mostly consists of carbon powder mixed with a PTFE emulsion, which is cured by a heat treatment. Clear directions for cost reduction have not been found. As the gas diffusion media play a critical role in the performance of the PEMFC, especially determining the maximum power output makes it worthwhile to focus on the GDM performance, rather than on the cost per  $\text{m}^2$ .

**Bipolar Plates** For bipolar plates, a total (bulk and contact) resistance value of  $10 \text{ m}\Omega \text{ cm}^2$  is usually specified [72]. This would result in  $0.02 \Omega \text{ cm}^2$  (two plates, coolant between), which at  $1.5 \text{ A cm}^{-2}$  would result in 30 mV voltage loss. Bipolar plates are made of carbon,

carbon/polymer composite, or metal. Both the resistance and weight/volume aspects demand thin bipolar plates. This has to be combined with low  $H_2$  permeability ( $<2 \times 10^{-6} \text{ cm}^3 \text{ cm}^{-2} \text{ s}^{-1}$ ) and good mechanical integrity.

From a materials durability point of view, carbon/polymer composite materials are to be preferred. However, metal-based bipolar plates enable the use of very thin plates, thus leading to an increase in volumetric power density. Major car manufacturers such as Honda and Toyota are using metal-based bipolar plates in the fuel cell stacks that are used in their latest generation fuel cell vehicles. Power density of  $1.9 \text{ kW L}^{-1}$  by both Honda [73] and Nissan [74] is ascribed to the use of metal-based bipolar plates.

From the total costs for metal bipolar plates as stated in the DoE cost review report, the costs of the metal plates amount to  $\$20 \text{ m}^{-2}$  active cell area, which will be at maximum 10% lower per  $\text{m}^2$  metal area. For the metal-based bipolar plates, the materials costs are significantly higher than the processing costs. As metal, coated 316L stainless steel is selected.

The use of metal-based bipolar plates brings two concerns [2]. At the anode, corrosion leads to the release of metal ions that can exchange with protons in the ionomer in the electrode as well as in the membrane, resulting in increased resistance of the electrode and membrane. At the cathode, the major concern is an increase in contact resistance, caused by the buildup of an oxide layer that has lower bulk conductivity than the metal itself. The stack compression force has an important effect on the contact resistance; the lower the compression force, the higher the contact resistance [75].

Various approaches are therefore being followed. A material, such as a stainless steel with high corrosion resistance, and with a low tendency to form a high resistance oxide layer, could be applied as plate material. Stainless steels need the right amount of chromium, nickel, and nitrogen to form a thin corrosion resistant layer on the surface that does not have a too negative effect on the contact resistance. Examples of stainless steels that have been qualified are 904L, and to a lesser extent 316L and 310S [76].

Recently, incorporation of nitrogen into the surface of especially nickel-based alloys by thermal nitration was shown feasible. In this way, the contact resistances

are reduced and the corrosion resistance of the materials is increased, thus meeting the targets set by the DoE [77]. Nitration of cheaper Fe-based stainless steels is being considered.

Another approach is to select steels that qualify when coated with a coating [2], that either prevents direct contact with the electrolyte and is not oxidized at the cathode. In this respect, ceramic metal nitride coatings are considered (Cr, Ti, TiAl nitride) [78]. If the coating is perfectly dense and stable, any base material could be applied. Alternative base metals being considered are aluminum or titanium [79].

The coating may be applied before or after the flow pattern is stamped in the plate, depending on whether it is possible to avoid damaging the coating layer during the forming process. Some stack manufactures advocate the use of thin gold coating, which results in a well-conducting coating but needs to exceed a thickness of several nanometers to be sufficiently dense and hence may not meet the cost targets [13]. Conductive polymers have also been suggested as coatings, although their stability under the fuel cell conditions is a concern [80].

Graphite/polymer composite materials made of a graphite, and carbon combined with a polymeric resin, are less susceptible to corrosion, and have low contact resistance but higher bulk resistance than metals. Also the gas permeability is higher than in metals. This results in thicker plates, which may still be acceptable for stationary applications. Target values for the bulk conductivity are in the order of  $>10 \text{ S cm}^{-1}$  [79].

Expanded graphite foil, as alternative to metal plates, would cost  $\$18 \text{ m}^{-2}$  [13, 81]. With respect to plate thickness and processibility [82], expanded graphite foil is similar to metal plates, enabling a stack power density of  $2.1 \text{ kW L}^{-1}$  [81].

Note that the flow field design of the bipolar plate is a critical factor for reactant supply and water management of stack and individual cells. As will be discussed later, a suboptimal flow field and manifold design can lead to an uneven distribution of reactants, and by this to fuel and air starvation.

**Seals and Edge Protection** The seals or gaskets provide protection to leakage of gases and coolants outside their compartments. They also control the stack height

and compensate for tolerances in material thickness. These materials consist of several components, the amount and type of which determines the final properties. Usually a base rubber is selected, to which additional fillers for mechanical strength and a plasticizer for flexibility are added. The seals have to be thermally stable as well as stable against hydrogen, oxygen, acid media, and coolants, and be compatible with other stack parts such as the bipolar plate.

Base rubbers may be based on silicone, fluoropolymers, or hydrocarbons. Although silicone rubbers such as silicone S and G have been applied in stacks, it has become clear that they are not sufficiently stable [83–85]. Materials like ethylene-propylene-diene-monomer (EPDM), butyl rubber (IRR), or fluororubbers (FKM such as Viton<sup>®</sup>) seem better suited. Further research is carried out to optimize properties like hardness, tensile strength, and stress relaxation. Also the morphology is being considered, with apparently a preference for profiled over flat gaskets.

Seals can be provided as separate stack components, but preferentially they are integrated in the MEA or the bipolar plate. This of course puts demands to the processibility of the base materials and the manufacturing process, which may involve curing to achieve cross-linking of the polymeric materials.

Misalignments of the anode and cathode, compression forces from the GDM, fibers stretching out from the GDM may all put electrochemical, chemical, and mechanical stress on the membrane, which the thin, state-of-the-art membranes cannot withstand. Therefore, some manufactures advise the inclusion of edge-protection on the membrane, that is, a layer of gas-tight and non-conducting material covering the membrane not covered by catalyst and GDL, and protruding slightly onto the area between the GDL (or catalyst layer) and membrane at the edges of the GDL. Suggested materials include polyimide [86]. These materials are there to reduce degradation of the membrane, but add to the complexity of the MEA and introduce yet other materials. Another option would be to let the edges of the electrodes coincide with the MEA, but this puts larger demands on the seals as well introduces the risk of shortening [13].

From the DoE cost review [13], it can be concluded that sealing of the cells comprises a higher cost than the

electrolytic membrane. Depending on the design, one needs one frame per MEA, or even three frames per MEA. The lack of literature on seals development is at least remarkable. Although sealing is a quite generic topic that plays a role in many devices where leakage of gases or liquids has to be prevented, the sealing in fuel cells is extra demanding, both due to the aggressive environment as well as due to the sensitivity of the PEMFC for contaminants.

## Operating Conditions Leading to Performance Loss and Shortening the PEMFC Lifetime

### Definitions and Targets

Besides cost reduction, durability is a key topic in the PEMFC R&D. The projected operational lifetime of a passenger vehicle is 5,000 h. While until recent years, PEM fuel cells R&D was especially focused on increasing the power density and decreasing the cost; attention during the last years has been shifted to durability issues. It has been recognized that during operation, the PEMFC cannot always be kept in its ideal window of operation, which has been found out to be steady operation at a constant voltage between 0.5 and 0.7 V, under fully humidified conditions and mild temperatures (70°C) using clean hydrogen and air [87]. Under these conditions, operating times far exceeding the required 5,000 h have been demonstrated with voltage decay rates of 1–2  $\mu\text{V h}^{-1}$ , which would limit the total voltage loss during a life of 5,000 h to 10 mV, which translates into a loss of efficiency of only 1%.

The numerous lifetime and durability studies on PEMFC in situ as well as on individual components ex situ have identified the main sources for performance decay and the related materials issues. A gradual decay of performance is mainly ascribed to effects in the cathode, notably loss of catalyst activity and increase of mass transport losses. The loss of activity was shown to result from dissolution and coarsening of the catalyst nanoparticles. The increased mass transport losses were ascribed to a decrease of the hydrophobicity of both catalyst layer and gas diffusion layer. Other factors contributing to the gradual decay are similar processes occurring in the anode and degradation of the bipolar plates and the seals of the MEA. The cause of sudden failure of the PEMFC was identified as the membrane losing its integrity, as a result of chemical as well as

mechanical stresses. These principal degradation mechanisms were found to be strongly dependent on conditions, where deviations from ideal conditions (70°C, 100% RH, constant load) usually exacerbate the effects. These deviations also include the presence of contaminants, either as degradation products from the system itself (bipolar plates, seals, catalyst ions, membrane fragments) as well as contaminants in the fuel, (CO, H<sub>2</sub>S, CH<sub>4</sub>, NH<sub>3</sub>) and in the air (SO<sub>x</sub>, NO<sub>x</sub>, NH<sub>3</sub>).

The targets currently accepted for automotive application are 5,000 h operation under drive cycle conditions or 40,000 h for stationary conditions with less than 10% performance decay. Such requirements imply that the decrease of mass activity should be limited and that the membranes will have to be able to perform without failure during the envisaged lifetime. The increase of mass transport losses should be minimal, putting demands on the stability of the support.

In this section, operating conditions and modes that contribute to voltage decay or limit performance will be discussed. In section “[Materials Degradation and the Relation to Performance Loss and Shortening the PEMFC Lifetime](#),” the durability issues related to the different components of the fuel cell, that is, catalyst, the gas diffusion layers, membranes, bipolar plates, and seals will be presented in more detail.

### Real-Life Conditions That Have an Impact on Fuel Cell Performance

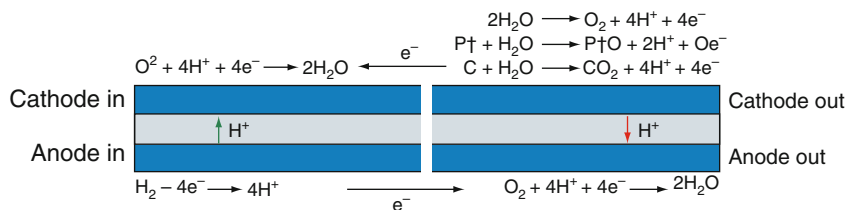
PEM fuel cells are in development to be ultimately deployed in the real world. The real world can vary considerably, from indoor use in buildings or warehouses to outdoor use under arctic conditions. Load variations can vary from continuous idle with occasional periods at maximum power for backup power systems, to highly dynamic operation in transport applications or working at continuous power for Combined Heat and Power systems. The level of control over all relevant parameters can be limited in a fuel cell system, and malfunctioning of control hardware can lead to conditions that fall outside the intended operating window. When compared to testing an MEA in a fuel cell laboratory in test hardware that excludes all influences but those deliberately applied by the operator, a large number of imperfections can be listed:

- The cell voltage is not constant during operation; extremes can be caused by systems malfunctioning
- The temperature and humidity control in the cell are limited, and are not constant; extremes can be caused by systems malfunctioning
- The feed of anode and cathode feeds can be fully interrupted due to systems malfunctioning, or become sub-stoichiometric in individual cells
- The inlet air will contain contaminants
- The hydrogen will contain contaminants, and for many applications other than road transport, only be a part of the anode feed
- Wear of cell and systems components may generate contaminants that can diffuse to other cell components susceptible to these contaminants
- The system might be placed in climate conditions, which are much harsher than anticipated in the design phase

### Cell Voltage Variations and Excursions

Especially in fuel cell vehicles, the load demand to the drive train is characterized by a very dynamic pattern. When viewed over the entire lifetime of the vehicle, it is most of the time shut off, in case of a passenger vehicle typically 95% of the time. During operation, the load demand to the drive train contains significant periods of idling, load cycles, and start-up/shut off periods. It has become apparent that especially those events leading to relatively high voltages, that is, low power demands, lead to durability problems.

In two recent papers by Nissan and Toyota [74, 88], degradation in fuel cells used in actual vehicles has been linked to the type of drive pattern and occurrence of non-steady state periods. In the paper by Nissan [74], it is shown that the correlation between performance loss on the one side and total vehicle operation hours and vehicle mileage on the other side is rather weak. It is especially the occurrence of start/stop cycles, idling and load cycling that contribute to the performance loss of the fuel cell stacks. Based on a model, verified with actual vehicle data, an estimate is made that 44% of performance loss can be attributed to start/stop cycles, 28% to idling, and 28% to load cycling. Vehicles used in Japan and the USA show different ratios, where start/stop cycles were occurring more frequently in Japan than in the USA, and the relative contribution of idling



### PEM Fuel Cell Materials: Costs, Performance and Durability. Figure 7

Electrochemical reactions take place when both hydrogen and air are present in the anode compartment (Figure previously published by the authors in Fuel Cells [87])

and load cycling is highly dependent on individual vehicle use.

During standstill, oxygen leaks in the anode compartment. In all cases where a hydrogen/air front arises, two horizontal fuel cell processes can take place as illustrated in Fig. 7. At open circuit, the electrons can only be exchanged horizontally in the same compartment, while protons will take the shortest distance in vertical direction through the membrane. In the bottom side, representing the anode, hydrogen is oxidized by oxygen. As the upper side, representing the cathode, lacks hydrogen, other oxidizable components can be oxidized, where the local potential is defined by the pull on electrons and protons by the oxygen reduction at both left and bottom of this cell section. This can lead to local potential values exceeding 1.5 V versus NHE. This mechanism has been previously described by Reiser [89]. The length of standstill appears to be an important factor as well: the longer the standstill period, the higher the oxygen concentration in the anode compartment becomes, and the more severe degradation becomes [74].

In the paper by Toyota [88], a marked difference is shown between performance loss and the type of simulated driving conditions. Whereas 20% performance loss is obtained after less than 100,000 km under low speed or congested conditions, it takes 350,000 km to reach this 20% performance loss under medium- to high-speed conditions. Analysis of the Japanese 10–15 drive cycle shows that 58% of the time the vehicle speed is zero or declining, that is, the fuel cell stack is operated at reduced power. Avoidance of high potentials leads to a sixfold decrease in the rate of performance loss: While operation between 0.65 V and OCV leads to 20% performance loss in less than 1,000 h, a lowering of 50 mV

of the maximum potential (OCV being equal to 1.0 V) leads roughly to a doubling of this time to reach 20% performance loss. However, it is clearly shown in the Toyota paper that this lowering of the upper voltage is at the expense of system efficiency, that is, better materials need to be developed to enable the high cell voltage excursions at periods of low power demand.

### Interruption of Anode and Cathode Feeds

Individual cells as well as complete stacks can be deprived of the fuel and air they need. In case of suboptimal stack design, the gas distribution over the individual cells becomes uneven, especially at high utilization. This can lead to individual cells running on sub-stoichiometric feeds, generally called fuel starvation when occurring at the anode, and air starvation when occurring on the cathode. When only individual cells are starved, these cells, placed in a series of hundreds of cells, are forced to generate the current that all other cells generate. When starved of fuel, other components present in the anode compartment are oxidized, starting those with the lowest Nernst potential and activation barrier. Carbon corrosion and platinum dissolution are two processes widely reported to take place under such conditions [90–92]. As water is relatively easy to oxidize as well, one of the mitigation strategies to protect the carbon and platinum is to include catalysts, such as iridium, that oxidize water at a lower potential than pure platinum. Knights et al. [90] have elegantly shown how the current generation requirement leads to a progressively increasing anode potential upon depleting the anode compartment of oxidizable species, up to the point that the anode potential surpasses that of the cathode, leading to cell reversal.

Air starvation is less well documented, and is believed to be less detrimental than fuel starvation. As the lack of oxygen needs to be replaced by other components to be reduced, this will generally lead to the reduction of protons to form hydrogen, which on itself leads to a negative cell voltage, as the proton reduction potential lies below the hydrogen oxidation potential [90].

One should be aware of the fact that fuel and air starvation lead to current-driven processes, with the resulting potential being an indication of the reaction taking over the normal fuel cell reaction. It is not an external potential imposed on the electrode that leads to an oxidation or reduction process.

### Shortage of Water: Temperature Overshoots and Concomitant Reduction of Relative Humidity

Cell temperature can differ from the design temperature structurally because of improper stack design, as well as occasionally due to malfunctioning of temperature probes or failing cooling systems. Over the cell area, there is always a limited temperature variation due to a finite heat transfer from cell to cooling medium. In air-cooled stacks, the temperature variations are generally bigger than when using liquid cooling media. Continuous higher temperatures of a few degrees than the design temperature will lead to a higher voltage decay rate, as many degradation mechanisms show a positive relation between degradation rate and temperature [87].

More severe in the context of PEMFC is the strong dependence of the relative humidity on the temperature. The delicacy of the water balance in the PEMFC has been described in section “[Introduction](#).” The water vapor pressure required for 100% relative humidity increases strongly with temperature. Extreme temperature excursions caused by malfunctioning of thermo couples or the cooling circuit, most probably, leads to drying of the MEA, which is difficult to restore in the fuel cell stack.

### Impact of Air Contaminants

In most fuel cell applications foreseen, a continuous flow of atmospheric air will be taken in while the system is in operation. This exposes the cathode, in the absence of air filters, directly to the substances

present in the air, their concentrations possibly diminished as a result of air humidification.

For transport, air contaminants such as CO, volatile organic components, nitrogen oxides, and sulfur dioxide are likely to be encountered, their concentration varying with the local air quality. Especially in the first period of introduction, where the fraction of zero emission vehicles is close to zero, while heavily polluting diesel engines are abundant, the fuel cell vehicle is likely to inhale dirty air. The mean value of typical contaminant concentrations in European cities are:  $42 \mu\text{g m}^{-3}$  for  $\text{NO}_2$  and  $12 \mu\text{g m}^{-3}$   $\text{SO}_2$  [93],  $1 \mu\text{g m}^{-3}$  is approximately 1 ppb. However, especially in Asia, much higher levels are common. In Shanghai, China,  $\text{SO}_2$  concentrations exceeding  $150 \mu\text{g m}^{-3}$  and  $\text{NO}_2$  concentrations exceeding  $170 \mu\text{g m}^{-3}$  have been measured [94]. In the end report of the HyFleet Cute project [95], a fuel cell buses demonstration project in eight European cities as well as Perth, Australia and Beijing, China, it is specifically mentioned that air contaminants in Beijing caused performance problems of the fuel cell stacks. In volcanic regions in Japan,  $\text{SO}_2$  concentrations as high as 20 ppm and  $\text{H}_2\text{S}$  concentrations as high as 4.5 ppm have been measured [96]. A drive cycle pattern developed by the Japanese FCCJ even includes a period for driving by a hot spring [97]. Some contaminants are very specific for certain applications of fuel cells, such as sulfur-containing combat gases as mustard gas for military use or salt-containing aerosols near the sea.

For rural areas, exposure to  $\text{NH}_3$  is likely to occur near intensive livestock farming. Ammonia concentration of 10 ppm is not uncommon, and was shown to lead to rapid drop in performance of around 10% in 2–4 h [98], of which 5% proved to be irreversible.

The influence of those contaminants on the MEA level has been assessed in a number of studies [96, 99] as well as reviewed in [100]. A linear dependence of the voltage drop on the  $\text{NO}_x$  concentration was observed at a current density of  $0.175 \text{ A cm}^{-2}$ , the above quoted  $42 \mu\text{g m}^{-3}$  would lead to a voltage drop of around 15 mV, while  $\text{NO}_2$  concentrations exceeding  $170 \mu\text{g m}^{-3}$  would lead to a voltage drop of 60 mV [99].

Narusawa et al. assessed the allowable concentrations of air contaminants on platinum cathodes [96]. While CO did not lead to any measurable poisoning at the cathode, presumably because the oxygen present

oxidizes CO at a high rate, NO<sub>2</sub> and SO<sub>2</sub> do lead to a loss in performance, albeit reversible. The allowable concentration of the air contaminants, defined as the concentration of a contaminant leading to a performance loss equal to 2 ppm of CO in the anode feed when using a Pt–Ru anode, is 257 ppm for CO, 2.6 ppm for NO<sub>2</sub>, and 1.8 ppm for SO<sub>2</sub>.

With respect to this reversibility, the data shown in [96] are not conclusive, as no graph demonstrates full recovery after exposure to NO<sub>2</sub> or SO<sub>2</sub> is stopped. In [98], exposure of a Pt cathode to 0.5 ppm SO<sub>2</sub> during consecutive periods of 2–4 h, leads during the first periods to a 5% decline in performance per period, without any intermediate recovery. Lower performance levels following next SO<sub>2</sub> periods were recoverable, but never to more than 90% of the original performance.

The simplest mitigation for air contaminants is the use of an air filter [101]. Activated carbon is widely applied as air filter for many applications requiring contaminant free air, as it has a quite generic adsorption capacity for both organic and inorganic contaminants. Although the application of an air filter comes at the cost of a pressure drop, which leads to extra compressor power, it is a relatively easy precaution that's worth considering. As the exposure to contaminants is unpredictable in the type and concentration of the contaminants, the breakthrough of the filter is unpredictable; frequent replacement of the air filter is probably the most practical and safe approach.

Mitigation of the influence of air contaminants by adapting the composition of the cathode catalyst is so far not applied.

### Impact of Hydrogen Contaminants

Technically, the hydrogen can be made as pure as necessary. For laboratory purposes, hydrogen quality is quoted as the percentage of hydrogen present in the gas. Hydrogen 6.0 stands for a gas containing 99.9999% of hydrogen, that is, the maximum of all other contaminants totals 0.0001%. For fuel cells, this quality standard is not helpful, as the standard does not discriminate between inert components and those with a detrimental effect on fuel cell performance.

Components that poison the fuel cell anode or membrane might have an adverse effect at even lower concentrations than 0.0001% (=1 ppmv, parts per

million by volume). Inert components can be tolerated toward a certain level, as their only effect is that of diluting the hydrogen. Although recycling hydrogen at the anode would lead to building up of such inert contaminants [102], it shouldn't lead to such strict purity requirements, as nitrogen cross-over from the cathode through the membrane takes place anyway. For the purpose of defining fuel cell grade hydrogen specifications, international actions are being pursued to identify the tolerable contaminants for PEM fuel cells [103]. Sixty components are suggested to have a potential harmful impact on fuel cell performance.

Much knowledge on the effect of poisons has been generated in connection to reformer-based PEMFC systems. Both for transport as well as for stationary applications, the presence of CO, CO<sub>2</sub>, NH<sub>3</sub> has to be taken into account besides N<sub>2</sub> and H<sub>2</sub>O [100]. For reformer-based systems that are operated dynamically, that is, including many cold starts and load variations, CO concentrations exceed the 10 ppm level frequently [104]. The effect of CO is studied most extensively. For unalloyed platinum electrodes, CO concentrations as low as 10 ppm lead to a performance loss of 100 mV [105] at 70°C. When reformer-based systems are fueled with logistic fuels, such as diesel and kerosene, other contaminants than CO and CO<sub>2</sub> are present in the reformat. Especially aromatics and unsaturated hydrocarbons can poison the fuel cell anode fast and irreversibly, even in concentrations so low that they are hardly detectable with state-of-the-art analytics.

Narusawa et al. assessed the allowable concentrations of hydrogen contaminants on platinum and platinum–ruthenium anodes [96]. Using the effect of 2 ppm CO on Pt–Ru anodes as benchmark, the allowable concentration of hydrogen contaminants is on Pt–Ru 0.33 ppm for HCHO (formaldehyde), 6.8 ppm for HCOOH (formic acid), 15 ppm for C<sub>6</sub>H<sub>6</sub> (benzene), 3.3 ppb for H<sub>2</sub>S (hydrogen sulfide), and 11 ppb for SO<sub>2</sub> (sulfur dioxide). On Pt anodes, these allowable concentrations are 0.05 ppm for CO, 0.56 ppm for HCHO, 17 ppm for HCOOH, 24 ppm for C<sub>6</sub>H<sub>6</sub>, 8.4 ppb for H<sub>2</sub>S and 10 ppb for SO<sub>2</sub>. Methane did not generate any effect at concentrations as high as 1,000 ppm.

The persistent issue of fuel cell poisoning in combination with fuel processor complexity has driven the vehicle manufacturers to pure hydrogen as the preferred fuel. For stationary and some off-road transport



applications, the use of gaseous and liquid carbon-based fuels is still preferred so that tolerance toward higher concentrations of species as CO is still pursued. The development of high temperature PEMFCs, such as phosphoric acid-doped PBI, that can be operated at 150–200°C, show a remarkable high tolerance toward CO. Concentrations as high as 3% CO can be tolerated at 200°C; at 125°C concentrations of 1,000 ppm, CO leads to only a minor voltage drop [106].

### Wear of Cell and Systems Components

Due to the sensitivity of the MEA components to fouling, wear of stack and systems components can have a negative influence on MEA performance. Best described is probably the effect of corrosion products when metal-based bipolar plates are used in stacks [2]. In addition, metal ions can be formed in other parts of the fuel cell system, upon contact with demineralized water used for cooling or humidification. The metal ions can exchange with protons on the electrolytic membrane, as well exchange the ionomer in the electrodes. As the conductivity of metal ions is much less than that of protons, this exchange effectively leads to a higher cell resistance.

Metal deposition on the electrocatalysts active sites can take place as well. Although elements such as chromium are used to enhance the activity of cathode catalysts, it must be in the alloyed form for such enhancement, while in the case of corrosion of metal plates, it will form an adlayer on the catalytic surface, blocking platinum atoms becoming inactive for oxygen or hydrogen dissociation.

Components leaching from seals are a potential cause for loss of MEA performance as well, although no such examples have been found. In long-term tests, seals have been noted to deteriorate completely, probably leading to the release of components as softeners and polymer fragments [87]. The direct influence of such components has not widely been reported.

### Freezing

Both transport systems, as well as outdoor installed stationary systems are exposed to climate conditions depending on their location. The vehicle DoE demonstration fleet, that is monitored by NREL, has been shown to be operated at ambient temperatures between  $-10^{\circ}\text{C}$  and  $+50^{\circ}\text{C}$  [107].

In most developed countries, freezing conditions occur during winter. The effect of freeze/thaw cycles has been studied to some extent. An immediate effect of freezing can occur during start-up. At  $-10^{\circ}\text{C}$ , the conductivity of Nafion, around  $0.025\text{ S cm}^{-1}$  [108], is enough to start the cell. As the water saturation pressure at this temperature is extremely low, liquid water will be formed and ice formation can easily occur, blocking the gas diffusion media pores, preventing start-up. It is among others for this reason that car manufactures have switched to metal bipolar plates, leading to a lower thermal mass of the fuel cell stack, and thus to a quick rise in temperature during start-up [109].

Without precautions, freeze/thaw cycles have a negative effect on cell performance, especially when liquid water in the MEA is still present under freezing conditions. Especially interfacial stresses can lead to delamination, while increasing ohmic resistances in membrane and electrode have been reported. The Daimler/EVO city buses demonstrated in the CUTE project were therefore always parked indoors overnight. In the latest systems, freezing of liquid water is avoided by water removal procedures when cooling down. Most passenger vehicles and buses of the major OEMs do not need indoor parking anymore.

Stationary PEMFC systems are even operated under arctic conditions, with a guaranteed lifetime of 4,000 h by Dantherm Power [110].

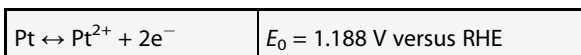
## Materials Degradation and the Relation to Performance Loss and Shortening the PEMFC Lifetime

### Catalyst Issues

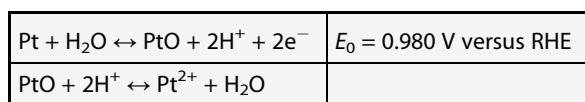
**Pt Coarsening/Dissolution** Pt catalysts in PEMFCs gradually lose their activity, mostly as a result of loss of ECSA. With present materials, the loss of ECSA can be in the order of 60% over ten thousands hours of constant load operation, and is found to be accelerated by load variations [87]. It has been well established that the loss of ECSA is due to growth of Pt particles as well as to loss of Pt into the ionomer phase of the MEA [111–113]. In a review by Shao-Horn et al. [114], four mechanisms are distinguished: (1) an electrochemically induced Ostwald ripening (2) electrochemical dissolution and deposition in the ionomer phase (3) migration

of crystallites on the carbon support followed by coalescence, and (4) detachment from the carbon support.

The Ostwald ripening and electrochemical dissolution are strongly related. Although Pt is one of the most stable elements in the Periodic Table, thermodynamics predict electrochemical Pt dissolution in acid media at the potentials as encountered at the cathode [115], either directly according to the reaction:



or indirectly through oxide formation



Electrochemical studies on Pt thin films, Pt wires, and Pt nanoparticles on carbon have shown that at potentials  $<0.85 \text{ V}$ , the dissolution process has a Nernstian potential dependence and that the equilibrium electrochemical potential of the Pt dissolution decreases with particle size [116]. Accordingly, higher equilibrium  $\text{Pt}^{2+}$  concentrations were measured for Pt/C than predicted for bulk Pt [111] at  $80^\circ\text{C}$ . According to mechanism 1,  $\text{Pt}^{2+}$  dissolved into the ionomer phase can redeposit on other (larger) particles in the cathode, due to the higher equilibrium potential of dissolution on these particles [117]. These particles need to be on the carbon support in order to allow for the exchange of electrons. The difference in dissolution potential is considered to be related to the surface tension. The above mechanism of dissolution and redeposition is an example of Ostwald ripening, that is, minimization of the surface energy is the driving force. The  $\text{Pt}^{2+}$  ions may also be reduced in the ionomer phase of electrode or in the membrane by  $\text{H}_2$  crossing over from the anode side (mechanism 2).

At potentials above  $0.85 \text{ V}$  versus RHE, it has been observed that Pt dissolution does not show the strong Nernstian potential dependence, and even has a maximum around  $1.15 \text{ V}$  [111, 118, 119]. This is ascribed to formation of the oxide layer, which impedes further Pt dissolution. However, upon lowering the potential, these oxides are reduced accompanied by Pt dissolution. It is thought that oxide formation is related to the observation that potential cycling enhances the

Pt dissolution. Kawahara et al. have studied the dissolution of Pt during cycling ex situ [120, 121]. It is known from literature of the early 1990s that fast cycling leads to oxides that are difficult to reduce, often called  $\beta$ -oxides [122, 123], and that the reduction of these  $\beta$ -oxides is accompanied by dissolution of platinum [124]. At potentials higher than  $1.2 \text{ V}$ , the formation of an oxygen skin was observed, with Pt atoms replacing oxygen atoms at even higher potentials leading to disintegration of the Pt surface [125].

Mechanism 3 and 4 are not electrochemical processes. For mechanism 3, Brownian motion is the assumed driving force, causing surface diffusion of particles with random collisions leading to coalescence [126]. Usually the fact that sintering does not occur significantly in catalysts in the gas phase at temperatures below  $500^\circ\text{C}$  is considered to be an indication that coalescence is not the prevailing mechanism [111, 127, 128]. Still, although both mechanisms 1 and 3 lead to an increase of the average particle size with an asymptotic particle size distribution (PSD), the coalescence mechanism has a log-normal distribution (tail at large sizes) and the Ostwald ripening has a tail at the smaller particle sizes but with a maximal particle size cut-off [129]. In many studies, a log-normal distribution of Pt particle sizes was found in virgin and used electrodes [111, 112, 127, 129, 130]. However, it must be noted that such analysis requires a good sampling also of small particles and results can be affected by the fact that several mechanisms are active at the same time [111, 127].

Accelerating factors are increasing temperature, potential cycling with the upper voltage limit being critical, and high relative humidity [130–132]. The effect of temperature on electrochemical dissolution is very strong, ex situ studies have shown an increase of orders of magnitude between  $65^\circ\text{C}$  and  $80^\circ\text{C}$  [118, 133]. It should be noted that reduction of the Pt loading, as required for cost reduction result, makes the system more sensitive for degradation to the extent that increased degradation rates were observed [132].

Non-nano-sized Pt is not immune for oxide formation or dissolution, but the equilibrium potential of oxide formation is higher compared to nano sized particles. Pt black and NSTF electrodes have either much larger Pt particles or even a continuous phase Pt. This results in lower specific surface area but more

stable activity as was convincingly demonstrated by the life-time studies carried out on NSTF electrodes by 3M [49].

**Stability of Alloys** Both at the anode and cathode, binary Pt–M alloys are proposed, either to enable operation on CO containing fuel or to increase the activity for the oxygen reduction reaction. The M–metal is usually a transition metal, that is, Co, Ni, Fe, Mn, Ir. These metals themselves are thermodynamically less stable than Pt. As regards the stability of the alloy, several factors play a role: the degree of alloying, the particle size, and the degree of ordering in the alloy.

When a Pt–M particle contains unalloyed M, this will easily dissolve at potentials of 0.6–0.9 V versus NHE. This has led to the use of pre-leached catalysts, some of which show improved stability even compared to pure Pt in this potential range [46]. However, recent work, which includes also cycling to potentials as low as 0 V versus NHE showed a strong decrease of activity. During this cycling, oxides that are formed at high potential and that have a passivating effect are reduced by cycling to low potential, followed by dissolution of the metal. On the other hand, the Pt seems to become less soluble [132].

The loss of the alloyed non-noble metal does not always lead to a decrease of activity. As already mentioned, in some cases, an increase of activity was found, ascribed to as surface roughening or electronic effects [61]. On the other hand, in other cases, the activity decreases because the beneficial effect on the electronic structure of Pt is lost. In any case, metal ions leaching from the catalyst will have a negative effect on the membrane and ionomer phase in the electrode.

Accelerating factors are similar to the factors accelerating the Pt coarsening and dissolution, that is, elevated temperature, potential cycling and high humidity, but for cathode catalysts, cycling to low potentials (i.e., 0 V vs NHE) seems to be an additional stressing factor [132].

Mitigating effects are also the same as the same as for Pt degradation. Large alloy particles are in general more stable than small particles [55], and at drier conditions, dissolution seems to be slower.

**Contaminants for Catalysts** Chemical species present in the fuel cell can further deactivate the catalyst,

either through surface poisoning or by (electro)chemical reactions, which drastically modify the catalyst. Such chemical species are then called contaminants, they can have their source outside as well as inside the system.

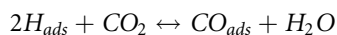
*At the anode:* CO adsorbs strongly on Pt at the fuel cell operating temperatures. Even at 10 ppm, the cell voltage loss is in the order of 100 mV. The effects are reversible and mitigated by increased temperature. Some Pt alloys mitigate the CO poisoning, either because the adsorption is less strong or because they facilitate the CO oxidation at low potentials. Among these alloys, well-dispersed Pt–Ru supported on carbon is widely accepted as the standard catalyst for reformat feeds. Tolerance is improved to 100 ppm, with a loss <20 mV at 500 mA cm<sup>-2</sup> at 80°C, using 0.5 mg cm<sup>-2</sup> Pt–Ru at the anode. These loadings seem high, but reformat application is only considered for stationary applications where such loadings may be acceptable. There is some debate on the mechanism of CO tolerance of Pt–Ru. For well-alloyed material, the interaction with CO is less strong, and on both alloyed and non-alloyed material, OH adsorption followed by CO oxidation can take place at low potential. Both mechanisms are probably involved, which one dominates will depend on the composition and preparation of the catalyst. The hydrogen oxidation on Pt–Ru is still sufficiently fast, even when just a part of the surface area is available.

Other catalysts, such as PtMo, mainly favor the CO oxidation. This is effective only if sufficient CO is oxidized at low potential, which is not always the case. Bilayer electrodes have been proposed to separate the CO oxidation from the H<sub>2</sub> oxidation. In this concept, CO oxidation takes place in a layer adjacent to the GDL, catalyzed, for example, by PtMo, and hydrogen oxidation in Pt–Ru layer adjacent to the membrane. This improved the CO tolerance significantly compared to the tolerance obtained with monolayer system of either catalyst [134, 135].

Dissolution of Ru or Mo from Pt alloys in anodes and migration to the cathode have been reported, and seem enhanced at high anode potentials, such as occurring during fuel starvation and severe poisoning [136]. PtWO<sub>x</sub> is expected to be much more stable, but the activity relies on the capacity to form bronzes, for which H<sub>2</sub> needs to dissociate on Pt. As the Pt becomes poisoned, these materials lose their activity [137].

Mixing small amounts (2–4%) of air into the fuel, the so-called air bleeds, can also be effective. The CO is then chemically oxidized. Air bleeds can have detrimental effect on durability, both on the catalyst as the chemical oxidation induces heat effects, as well as on the membrane as detrimental radicals such as OH• can be formed.

Reduction of the CO content in reformat feeds below 10 ppm is usually not advantageous, since the resulting feed still contains high fractions (10–25%) of CO<sub>2</sub>. On Pt in the potential range where adsorbed H is present, CO<sub>2</sub> reacts along the so-called reverse Water Gas Shift reaction:



The reaction product is adsorbed CO, poisoning the catalyst. Again this effect is mitigated on Pt–Ru due to the strong adsorption [138–140].

The tolerance for H<sub>2</sub>S is usually <0.1 ppm. H<sub>2</sub>S will decompose on the anode surface and poison the Pt, an effect, which can be irreversible at high potential or with high S concentration when Pt–S is formed (thermodynamically at very low potential already). H<sub>2</sub>S can migrate to the cathode and form Pt–S there as well. The only way to remove Pt–S is to oxidize it at very high potential (1.2 V vs NHE). Pt-based catalysts promoting S oxidation at low temperature are not yet known [100].

A remarkable finding in the previously cited study by Narusawa [96] is that while alloying of platinum with ruthenium mitigates the poisoning of CO, it aggravates the poisoning of many other contaminants, such as HCHO (formaldehyde), HCOOH (formic acid), H<sub>2</sub>S (hydrogen sulfide), and C<sub>6</sub>H<sub>6</sub> (benzene). For all contaminants studied, the addition of an alloying metal does not make irreversible adsorption reversible, that is, the decrease of the potential at which oxygen and hydroxyl adsorption start is not high enough to shift the onset of the oxidation of adsorption to a useful potential range.

*At the cathode:* Contaminants with a strong interaction with the cathode catalyst might accumulate over time and lead to performance loss after reaching a certain threshold.

When a fuel cell is operated during 5,000 h at an air stoichiometry of 2, the accumulated number of contaminant molecules present in a concentration of

1 ppmv outnumber the amount of platinum surface atoms by a factor of 300. Molecules that adsorb irreversibly on platinum will easily poison the platinum surface to such an extent that fuel cell performance is bound to approach zero in the course of 5,000 h. While the surface coverage of the contaminant slowly builds up, the available sites for oxygen reduction decrease. As two adjacent platinum sites are needed for oxygen desorption, the oxygen reduction activity at constant potential is proportional to  $(1 - \theta_c)^2$ , in which  $\theta_c$  stands for the degree of coverage by the contaminant.

Except for ammonia, all air contaminants are believed to lead to poisoning of the catalytic sites. Ammonia is thought to lead to a loss of ionic conductivity in the ionomer in the cathode layer, as a result of an acid–base reaction between the basic ammonia and the acidic sulfonic acid group of the ionomer.

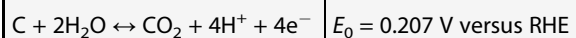
SO<sub>2</sub> in air can chemically adsorb on the catalyst, but also reduction of SO<sub>2</sub> on Pt leads also to Pt–S formation, with Pt–SO as intermediate product. Surface adsorption of NO on Pt has also been reported. The effect of NO<sub>2</sub>, on the other hand, is not related to surface adsorption, rather NH<sub>4</sub><sup>+</sup> is formed by electrochemical reduction, which poisons the membrane conducting phase (see below) [100, 141].

*Internal sources:* Some metal ions may deposit on the catalyst, that is, such as Ru on Pt cathode [136], which will deactivate them. Also Cl<sup>−</sup>, which may be present as a result of the catalyst manufacturing is a strong poisoning for Pt catalysts [142].

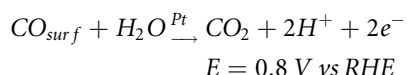
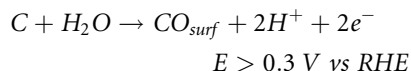
**Carbon Corrosion** Both electrochemical oxidation and thermal degradation of carbon in humid air at temperatures <125°C have been reported, and it seems established that these corrosion mechanisms are accelerated by the presence of Pt. Carbon corrosion will first modify the surface of the support, which will become less hydrophobic. It has also been reported that carbon corrosion may enhance the mobility of Pt on the surface, accelerating the Pt sintering discussed above. Further carbon corrosion will degrade the electron-conducting network, rendering Pt particles inactive.

Thermodynamically carbon is less stable in acid media than, for example, Pt, but due to the slow reaction kinetics, carbon can still be used in fuel cell [143].

Carbon (graphite) can be electrochemically oxidized to CO<sub>2</sub> at quite low potentials:



Electrochemical ex situ studies [131, 144, 145] in the temperature range 25–80°C have shown that at a potential higher than 0.3 V versus RHE, CO<sub>surf</sub> starts to form irreversibly on the carbon particle surface. One specific species is the quinone group that is electrochemically active with a redox peak at 0.55 V versus RHE that can be identified in cyclic voltammetry. The presence of Pt catalyzes the subsequent oxidation to CO<sub>2</sub>. The carbon corrosion mechanism consists of the following steps:



In the absence of Pt, CO<sub>2</sub> emission was observed only at potentials above 1.1 V versus RHE. Carbon corrosion is often quantified in potential hold tests, where the catalyst is held at potentials >1.2 V [145, 146].

The thermal degradation of carbon and platinum-loaded carbon in air is not expected to take place below 100°C [147], although at higher temperatures, it was shown to be accelerated by the presence of Pt nanoparticles. However, the humidification of air substantially enhances the thermal corrosion rate of carbon, by providing an additional pathway for chemical carbon oxidation through a direct reaction with water [148, 149].

In most carbon corrosion studies, it is observed that the corrosion rate (in g h<sup>-1</sup>) increases with the specific surface area (m<sup>2</sup> g<sup>-1</sup>). High surface area carbons have more edge features and are therefore sensitive to oxidation centers [144]. An effective method to reduce the carbon corrosion is therefore to decrease the number of dangling bonds, that is, use a more graphitic carbon. Heat-treated carbon was shown to be more stable as were materials that by nature have a more graphitic surface such as carbon nanotubes, as well as some type of carbon nanofibers. Good stability has been shown in

fuel cell accelerated stress tests but a drawback is the lower surface area of these materials, which results in larger metal particles or lower metal loading [150, 151].

Accelerating conditions for carbon corrosion include high potentials as during air/hydrogen fronts and fuel starvation. The role of humidity is still under some dispute. On the one hand, the carbon corrosion reactions require the presence of water; on the other hand, oxidation of water may proceed at higher rate than carbon corrosion. The recent research into water oxidation catalysts to mitigate carbon corrosion is based on this. However, it must be noted that the inclusion of such (metal-alloyed) catalysts has concomitant durability issues.

### Gas Diffusion and Microporous Layers

The GDL and MPL experience conditions much similar to the catalyst layer, only there is no ionomer to provide protons, and there is no Pt catalyst to enhance reactions. The water phase is acidic due to the presence of degradation products from other components (CO<sub>2</sub>, SO<sub>3</sub><sup>-</sup>, F<sup>-</sup>), resulting in a pH of about 4 [152]. The presence of fluorinated binders in MPL and GDL protects the carbon to some extent, but surface oxidation or even oxidation to CO or CO<sub>2</sub> can occur in the environment of liquid water, with O<sub>2</sub> present in the gas phase and dissolved in water. Schulze et al. present evidence for decomposition of PTFE on the basis of XPS data, but a mechanism has not been proposed [133].

The result of these degradation mechanisms is that that the GDL and the MPL both lose their hydrophobic character [133, 153, 154], and that the pore structure of the materials changes. The relation between microstructure and surface properties and mass transport properties has been the subject of several recent experimental studies [155, 156], which indicate that indeed mass transport can be seriously affected by the hydrophobicity of the GDL and MPL as well as by the pore size. This will contribute to the gradual decay of the performance, though it is hard to distinguish the effects of changes in the GDL/MPL to those of changes in the catalyst layer.

The decreased hydrophobicity can result in widening the window in which flooding can occur. Flooding itself is not regarded as a main cause for degradation.

Although in some overviews it is stated that flooding might lead to accelerated carbon corrosion and platinum dissolution, one should realize that the occurrence of both flooding and a high cathode potential is unlikely. The potentials needed to lead to significant platinum dissolution and carbon corrosion are 0.9 V and higher [87, 157], hardly a potential that occurs at the cathode during high power output.

An important function of the GDL is to even out the compression forces in the stack. These compression forces are also a source for degradation. Lee and Mérida concluded from *ex situ* experiments that the compressive strain increased with the applied pressure but even more strongly with temperature, and the GDL strain was influenced by the PTFE stability [158]. Properties such as in-plane electrical resistivity, surface contact angle, bending stiffness, and porosity were not affected. However, it was found that convective airflow through the GDL under strain can lead to loss of material. The GDL degradation can also contribute to sudden failure as GDL fibers can puncture the membrane, either when as a result of degradation they do not distribute the compression forces well any more or when they are even broken themselves.

## Membranes

**Chemical Stability** Chemical instability of membranes has always been associated with the formation of peroxide in the PEMFC. Peroxide and related radicals can attack aliphatic groups, resulting in chemical degradation. Therefore, perfluorinated or highly aromatic polymers have been considered only, with perfluorinated materials now being the commercially available option. The removal of end-groups like carboxyl or H has resulted in a much enhanced stability. An important recent development has been the identification of the hydroxyl ( $\bullet\text{OH}$ ) radical as the detrimental species rather than peroxide [159]. For this radical to form from  $\text{H}_2\text{O}_2$ , the presence of oxidizable metal ions such as  $\text{Fe}^{2+}$  is required; hence, it was thought that contaminating ions were required to induce radical formation. Recently, however, it was reported [160, 161] that they can be formed under fuel cell operation in a direct reaction of  $\text{H}_2$  and  $\text{O}_2$ . It is now assumed that  $\text{H}_2\text{O}_2$ ,  $\bullet\text{OH}$ , and  $\bullet\text{OOH}$  are formed due to gas cross-over reaction of  $\text{H}_2$  and  $\text{O}_2$  either chemically at

electrodes and in the membrane or electrochemically at the anode, and that their concentrations are related through the equilibrium:

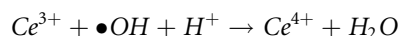


The presence of these radicals would lead to the classical unzipping mechanisms induced by an attack on C–H and COOH groups, present in small quantities in the polymer after manufacturing. This mechanism would result in continuous HF release, no chain scission, and “ideal PFSA,” not containing such reactive end-groups, would not be vulnerable [162].

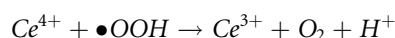
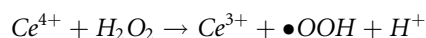
Still, also improved PFSA materials suffer from chemical degradation as was established by increasing HF release, which indicates a chain scission process [159]. Further analysis showed that the  $\bullet\text{OH}$  radical can react with  $\text{H}_2$  to form  $\text{H}\bullet$ , which can react with the tertiary fluorine atom, thus initiating a main chain scission process, as shown in Fig. 8.

A second mechanism involving  $\bullet\text{OH}$  is exacerbated by dry conditions, when the deprotonation of the sulfon groups is incomplete. The proton on the sulfon group reacts with the  $\bullet\text{OH}$  radical, leaving a sulfonyl radical, which has a weak bond to the carbon atom. This initiates another scission mechanism. This is illustrated in Fig. 9.

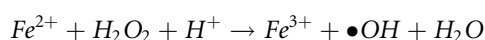
Recognition of these two mechanisms has led to the conclusion that  $\bullet\text{OH}$  is the “killer species” [159]. A possible mitigation strategy is the addition of radical scavengers, such as  $\text{Ce}^{3+}$  or  $\text{Mn}^{3+}$ , which react with  $\bullet\text{OH}$



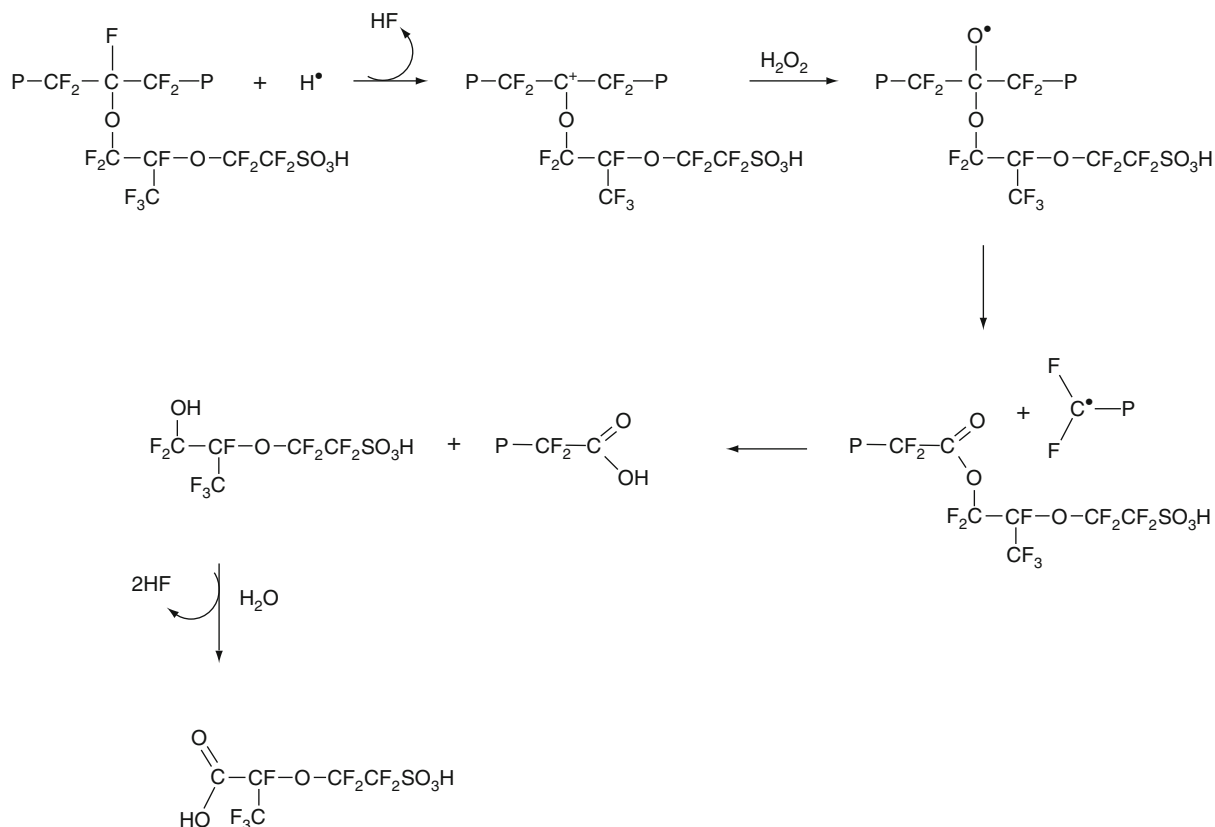
and which are recovered through reaction with  $\text{H}_2\text{O}_2$ ,  $\bullet\text{OOH}$ ,  $\text{H}_2\text{O}$ ,  $\text{H}_2$



Unlike  $\text{Fe}^{2+}$  ions, the mitigating ions should *not* reduce  $\text{H}^+/\text{H}_2\text{O}_2$  as this would result in  $\bullet\text{OH}$  formation, as is the basis of Fenton’s test [87].

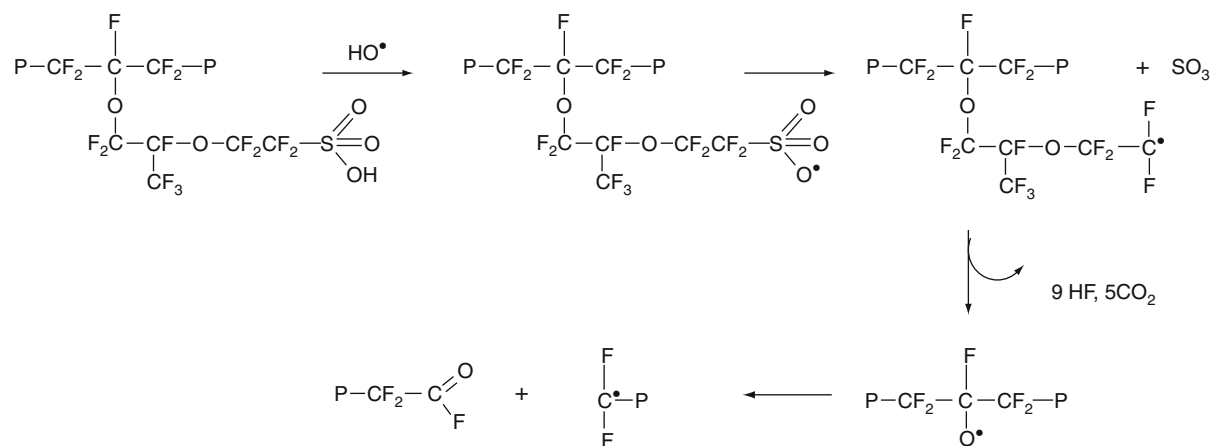


This requirement involves careful matching of reduction potentials and reaction rates but has led to much increased stability [163–165].



PEM Fuel Cell Materials: Costs, Performance and Durability. Figure 8

Scheme showing the degradation of "ideal" PFSA induced by the attack of a  $\text{H}^\bullet$  radical on a tertiary fluorine atom (After reference [159])



PEM Fuel Cell Materials: Costs, Performance and Durability. Figure 9

Scheme showing the degradation of PFSA in dry conditions initiated by an  $\bullet\text{OH}$  radical (After reference [159])

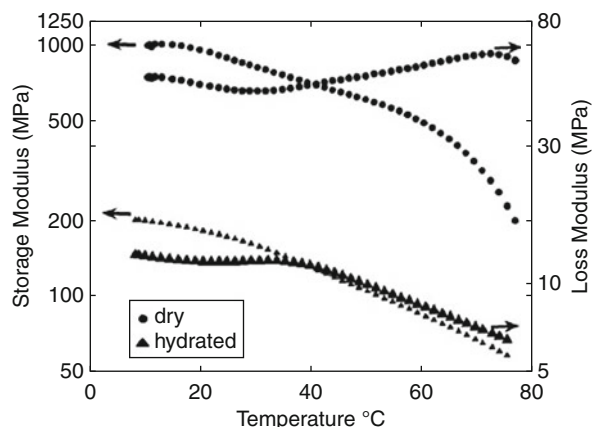
The relation between the chemical degradation of the membrane and cross-over of the reactant gases has also shed new light on the stability of (aromatic) hydrocarbon-based membranes such as sulfonated PEEK. Although such materials perform dramatically bad in the Fenton's test, which shows that they are vulnerable for attack by  $\bullet\text{OOH}$  and  $\bullet\text{OH}$ , stable performance has been obtained for such membranes [28, 166]. This may be associated with the gas cross-over rates that are lower in hydrocarbon membranes than in PFSA membranes by a factor 5–40 [28, 167], thus reducing the rate of radical formation. However, depending on the degree of aromaticity of some non-fluorinated compounds such as PEEK are susceptible to the loss of the sulfonic acid group by hydrolysis at wet conditions [166], while others like polysulfones are much more stable [168].

Chemical degradation of PFSA membranes seems to be accelerated by elevated temperature, OCV conditions, low relative humidity, and elevated gas pressures. The main mitigation is the addition of radical scavengers and removal of non-perfluorinated groups. Note, that the amount of metal scavenger ions added to the membrane is limited and does not affect much the conductivity.

**Mechanical Degradation** Sulfonated membranes take up liquid water, which enables them to conduct protons, but which also results in swelling of the membrane. Hence, stresses occurring during temperature and humidity cycling [169, 170] can lead to mechanical failure of the membrane. Mechanical failure can also result from nonuniform contact pressure [171], high differential initial gas pressure over the membrane, and punctures as well as fatigue from stresses occurring.

Fig. 10 shows the effect of both temperature and hydration on the storage and loss modulus of Nafion NRE212, as measured by dynamic mechanical analysis [172]. While temperature has already a marked effect on the elasticity of Nafion, as expressed by the drop of the storage modulus, hydration leads to a decrease in the storage modulus over the whole temperature range.

Therefore, most commercial membranes are now reinforced, for example, by a porous polyethylene or PTFE [170, 173] material that enhances the dimensional stability and reduces the shrinkage stress in the membrane during drying. Reinforced membranes



**PEM Fuel Cell Materials: Costs, Performance and Durability. Figure 10**

Storage modulus and loss factor as a function of temperature for dry and hydrated as-received Nafion NRE212 (Reprinted from [172]; with permission from Elsevier)

show a lower decrease of the OCV [169] and longer lifetime at elevated temperature and low humidity [173].

A successful way to achieve good proton conduction at low humidity has been through low EW PFSA. A durability problem is that these low EW materials suffer loss of crystallinity and therefore are also easily water soluble, resulting in a lack of mechanical integrity, which shows, for example, during humidity cycling. Several solutions have been suggested such as cross linking, either of the backbone or of the sulfonic groups, blending and changing the protogenic group as was discussed in section “PEMFC Component Costs and Performance: Targets, Status and Developments”.

**Freezing Conditions** Freezing conditions result in a reduction of the conductivity of the PFSA membranes but do not seem to induce irreversible changes [174]. In a PFSA membrane, 90% of the water can freeze, but some water is non-freezable. The temperature at which freezing is observed varied with the water content. McDonalds et al. studied N112 membranes that underwent up to 385 F/T cycles between  $-40^{\circ}\text{C}$  and  $+80^{\circ}\text{C}$ . These materials were in a relatively dry state (only humidified by ambient conditions). These studies did not indicate any serious physical or chemical



damage. The toughness of membranes seemed to decrease somewhat, as well as the permeability of the membrane for oxygen. It was suggested that a rearrangement of the sulfonic acid groups in the ionomer might have occurred.

**Contamination by Ionic Species** Metal ions that originate from other parts of the system (electrode, bipolar plates, coolant, BOP) or from the fuel or air can be absorbed into the membrane.  $\text{NH}_3$  present in the fuel or in air will migrate into the membrane to form  $\text{NH}_4^+$ . Another source of  $\text{NH}_4^+$  is  $\text{NO}_2$ , that can be present in air and can be electrochemically reduced at the cathode [141]. The sulfonate sites have a stronger affinity to cations except  $\text{Li}^+$ , than for protons, leading to exchange of the protons by the metal ion when present. The direct effect is a reduction of the proton conductivity that is proportional to the decrease in proton concentration [171, 175]. An indirect effect is the lower hydrophilicity of metal ions in comparison to protons, which can result in partial drying out of membrane, thus also reducing the proton conductivity.

### Bipolar Plates

Bipolar plates are made from graphite, graphite/polymer composite, or metal. A loss of mechanical integrity of the bipolar plate would result in mixing of the fuel or oxidant with the coolant, or even of mixing fuel with air, leading to severe degradation of performance or failure. Moreover, the electronic current is passed from one cell to the other by the polar plate, which means that any change in intrinsic resistance or contact resistance will affect the ohmic losses. The conditions experienced by the bipolar plates are similar to the ones experienced by the GDLs, with the addition of exposure to the coolant.

For graphite and graphite composite plates, corrosion and release of contaminants are under normal operation not an issue. At normal operating conditions, the cathode potential is not high enough for oxide formation on the surface, and the contact resistance remains constant during fuel cell operation [75, 176]. Especially under start/stop or fuel starvation conditions, electrodes can be exposed to higher potentials but the pH is about 4 [152], that is, much higher than the catalytic layer. However, chemical degradation

is possible by oxidation of carbon in a reaction with  $\text{H}_2\text{O}_2$ , which may be formed due to  $\text{O}_2$  gas crossing over to the anode during a stop period [177]. The polymer of the composite plate may be subject to similar degradation effects, which may in the end lead to embrittlement, leaching from the coolant or gas cross-over. A serious durability issue with graphite/composite plates is the prolonged exposure to temperatures in the order of  $120^\circ\text{C}$ , which may result in creep and deformation of some resins, or the brittleness of thermosetting resins.

The use of metal bipolar plates brings about serious durability issues, as was already discussed in section “Bipolar Plates” Most metals form a passive layer on the surface at the potential in which the cathode operates, which affects the contact resistance. As discussed, this requires either a dense, stable, and conductive coating or a surface modification that enhances contact resistances as well as improves stability. At the anode, the conditions are usually not such that oxide layers are formed, but during starvation, poisoning or start/stop oxidation is accelerated resulting in metal ions being directly released into the fuel cell. Similarly, low potentials occurring at the cathode may result in reduction of the passivating layer, thus releasing metal ions. Such ions will further contaminate membrane and catalysts.

### Seals Degradation [87]

Fuel cell stacks contain seals on the MEA side as well as on the coolant side of the bipolar plate. Not only are they meant to prevent leakage of the gases and the coolants outside their containments, seals function as electrical insulation, stack height control, and variability control as well [178]. The degradation phenomena connected to seals do not only refer to the loss of the functionality of the seals themselves, but also to the leakage of seal components that could poison the MEA.

Degradation phenomena on seals are in general poorly understood, and only a very few number of papers have been published on seal degradation [84, 85, 178, 179]. Papers from Frisch [178] and Du [179] report on the selection of seals that meet fuel cell requirements. PlugPower used an ex situ method to test the seals resistance against a specific coolant as well as against acids that mimic the fuel cell

environment [179]. Weight change and the release of contaminants are used as indicators for the compatibility of the seals. Without giving hints to specific materials, the differences between various materials appeared to be huge.

Dow Corning [178] considered silicone elastomers as seals. The stability against coolant is not seen as a problem by Dow Corning for these silicone elastomers, while there is concern about the stability in the MEA compartment. Exposure of sealing materials to a mixture of various acids in a 1 M concentration at undefined temperature is used as *ex situ* aging test. General purpose silicone elastomers show cracks after 336 h in such a test, while special fuel cell grade silicone elastomers survive such a test for over a year without showing cracks.

In a study by Tan [85], the stability of various sealing materials has been tested in simulated environment at 60°C and 80°C. This simulated environment consisted of solutions containing HF and H<sub>2</sub>SO<sub>4</sub>, in two different concentrations. It was concluded that Silicone S and Silicone G are heavily degraded in the concentrated solutions as well as the diluted concentrations, although most of the data are collected in the concentrated solutions. Degradation reveals itself by weight loss, complete disintegration, as well as by leaching of Mg and Ca. The latter stem from magnesium oxide and calcium carbonate, which are used as fillers for obtaining the desired tensile strength, hardness, and resistance to compression. When these components are leached out, mechanical properties will be lost, and one might expect a negative influence on fuel cell performance as well, as these components can replace protons in the membrane as well as affect the properties of gas diffusion media and electrodes. An increase in temperature, as well as exposure to stress accelerates the degradation. The degradation mechanism is thought to involve de-crosslinking as well as backbone scission. Materials that much better survive the exposure to the solutions are ethylene-propylene-diene-monomer (EPDM) and fluoroelastomers.

Only in a couple of long-term experiments was seal degradation observed, and this might have been the consequence of an inappropriate materials selection. Silicone seals in direct contact with a perfluorosulfonic acid membrane suffer from degradation, at the anode as well as at the cathode [84].

The degradation is probably caused by acidic decomposition of the sealing material, leading to coloration of the membrane and detectable amounts of silicon on the electrodes. No fuel cell performance loss or increase in gas leakage along the seal has been observed. The same observation has been made by St. Pierre et al. after an 11,000 h test with a Ballard Mark 513 stack [180]. The seals were visibly oxidized, albeit more in the humidification section than in the active section. In the 26,000 h testing by Gore, complete degradation of the glass-reinforced silicone seal has been observed [14]. It forced them to increase the compression force to keep the cell gas tight. This might have had an impact on effective porosity of the gas diffusion media. The silicon of the seal could be detected throughout the MEA, especially on the gas diffusion media.

### Future Directions

Fuel cell vehicles are being demonstrated worldwide, and have reached a maturity that allows daily driving in demonstration schemes while offering a performance that meets consumer expectations. The latest generation fuel cell vehicles combine the user characteristics of modern passenger cars, such as a maximum speed of around 160 km h<sup>-1</sup>, a driving range of 400 km or more, with a low energy use. Refueling can be done in a couple of minutes. The progress in materials development from the past years has brought the cost, performance, and durability targets set for market introduction of road transport applications in sight. The *caveat* with respect to costs as presented in this contribution is that they are based on an annual production volume of 500,000 vehicles. The fuel cell systems in the first generation of commercially produced vehicles will certainly be more expensive than those projected for the 500,000 vehicles volume.

As cost reduction and durability improvement will not automatically go hand in hand, it is important to define priorities. For the acceptance of a new technology, robustness and performance are crucial, more than costs. With many alternative drive trains and energy carriers foreseen to be introduced in the coming decade, failures on robustness and performance will not be tolerated by consumers. If market stimulation packages by governments to stimulate the sales of clean vehicles are put in place, the last part of the cost

reduction can be postponed to next generation vehicles, when the technology already has been adopted. In any case, the options for further cost reduction should be present.

For cost reduction, the major contribution has to come from alternative electrodes, in which the mass activity of platinum is driven to its maximum, and transport limitations are driven to their minimum. A better control on catalyst layer and gas diffusion medium structures will be crucial in this respect, both on design and on maintaining the beginning-of-life properties. In this respect, the focus should be directed more than in the past on materials that are intrinsically stable under the harsh fuel cell conditions. Theoretically, the targets are achievable with presently known materials.

For robustness, it seems of key importance to use proton-conducting materials that can be used at higher temperatures and lower relative humidities than presently used materials. Application of such materials in the membrane as well as in the electrodes can lead to a considerably simpler and more robust fuel cell system. The combination of new components with other new or already existing components that need to be integrated in a new MEA is part of the development work that should not be underestimated. Given the benefits that fuel cells offer for clean and efficient transport, it will be worth the effort.

## Bibliography

### Primary Literature

1. US Department of Energy (2007) Multi-year research, development and demonstration plan, hydrogen, fuel cells & infrastructure technologies program, DOE/GO-102007-2430
2. de Bruijn FA, Makkus RC, Mallant RKAM, Janssen GJM (2007) Materials for state-of-the-art PEM fuel cells, and their suitability for operation above 100°C. In: Zhao T, Kreuer KD, Nguyen T (eds) *Advances in fuel cells*. Elsevier, The Netherlands
3. Zawodzinski TA Jr, Derouin C, Radzinski S, Sherman RJ, Smith VT, Springer TE, Gottesfeld S (1993) Water uptake by and transport through Nafion® 117 membranes. *J Electrochem Soc* 140:1041–1047
4. Gasteiger HA, Panels JE, Yan SG (2004) Dependence of PEM fuel cell performance on catalyst loading. *J Power Sources* 127:162–171
5. TIAX LLC (2003) Platinum availability and economics for PEMFC commercialization, DOE report number: DE-FC04-01AL67601
6. Ralph TR, Hogarth MP (2002) Catalysis for low temperature fuel cell Part I. The cathode challenges. *Platinum Met Rev* 46:3–14
7. Heinzel A, Mahlendorf F, Niemzig O, Kreuz C (2004) Injection moulded low cost bipolar plates for PEM fuel cells. *J Power Sources* 131:35–40
8. Stumper J, Stone C (2008) Recent advances in fuel cell technology at Ballard. *J Power Sources* 176:468–476
9. Janssen GJM (2001) A phenomenological model of water transport in a proton-exchange-membrane fuel cell. *J Electrochem Soc* 148:A1313–A1323
10. Weber AZ, Newman J (2005) Effects of microporous layers in polymer electrolyte fuel cells. *J Electrochem Soc* 152:A677–A688
11. Hurvitz N (2008) An in-situ, real-time gas humidity sensor for fuel cells, fuel cells durability and performance. The Knowledge Press, Brookline, pp 231–244
12. Satayapal S (2009) Overview of hydrogen and fuel cell activities, 27-10-2009 Fuel Cells & Hydrogen Joint Undertaking Stakeholders General Assembly, Brussels
13. Ernst WD, Stone C, Wheeler D (2009) Fuel cell system cost for transportation-2008 Cost Estimate, NREL/BK-6A1-45457
14. Cleghorn SJC, Mayfield DK, Moore DA, Moore JC, Rusch G, Sherman TW, Sisofo NT, Beuscher U (2006) A polymer electrolyte fuel cell life test: 3 years of continuous operation. *J Power Sources* 158:446–454
15. Yamazaki O, Oomori Y, Shintaku H, Tabata T (2005) Evaluation Study of PEFC Single Cell at Osaka Gas, 2005 Fuel Cell Seminar Abstracts. Courtesy Associates, Washington
16. Huth H (2008) Volkswagen's high temperature polymer electrolyte fuel cell, 4th annual international conference fuel cells durability and performance, Cambridge, 9-12-2008
17. Perti D (2009) DuPont next generation membrane and membrane electrode assembly development. In: FC Expo 2009, Tokyo
18. Johnson WB, Bazkowski C, Berta T, Crum M, Greene L, Kunitz B, Mao H, Priester S, Rudolph J, Ryan K, Seligura C (2011) MEA degradation issues opportunities and challenges using thin, reinforced polymer electrolyte membranes. In: 2nd international workshop on degradation issues on fuel cells, Greece, 21–23 Sept 2011
19. Hicks MT (2006) MEA and stack durability for PEM fuel cells. DOE hydrogen program FY 2006 annual progress report, pp 722–726
20. Sone Y, Ekdunge P, Simonsson D (1996) Proton conductivity of Nafion 117 as measured by a four-electrode AC impedance method. *J Electrochem Soc* 143:1254–1259
21. Maalouf M, Pyle B, Sun CN, Wu D, Paddison SJ, Schaberg M, Emery M, Lochhaas KH, Hamrock SJ, Ghassemi H, Zawodzinski TA (2009) Proton exchange membranes for high temperature fuel cells: equivalent weight and end group effects on conductivity. *ECS Trans* 25:1473–1481

22. Aquivion PFSA mebrane performance data. [http://www.solvaysites.com/sites/solvayplastics/EN/specialty\\_polymers/Specialties/Pages/Aquivion\\_PFSA.aspx](http://www.solvaysites.com/sites/solvayplastics/EN/specialty_polymers/Specialties/Pages/Aquivion_PFSA.aspx)
23. Hamrock S (2009) Membranes and MEAs for dry, hot operating conditions. DOE hydrogen programme FY 2009 annual progress report, pp 1042–1047
24. Cleghorn S, Griffith M, Liu W, Pires J, Kolde J (2007) Gore's development path to a commercial automotive membrane electrode assembly. 2007 fuel cell seminar. Courtesy Associates, Washington
25. Zhang YM, Li L, Tang J, Bauer B, Zhang W, Gao HR, Taillades-Jacquin M, Jones DJ, Roziere J, Lebedeva N, Mallant R (2009) Development of covalently cross-linked and composite perfluorosulfonic acid membranes. *ECS Trans* 25:1469–1472
26. Jones DJ, Rozière J (2003) Inorganic/organic composite membranes. In: Vielstich W, Gasteiger HA, Lamm A (eds) Handbook of fuel cells-fundamentals, technology and applications, vol 3. John Wiley & Sons, Chichester, pp 447–455
27. Kerres J (2005) Blended and cross-linked Ionomer membranes for application in membrane fuel cells. *Fuel Cells* 5:230–240
28. Aoki M, Asano N, Miyatake K, Uchida H, Watanabe M (2006) Durability of sulfonated polyimide membrane evaluated by long-term polymer electrolyte fuel cell operation. *J Electrochem Soc* 153:A1154–A1158
29. de Araujo CC, Kreuer KD, Schuster M, Portale G, Mendil-Jakani H, Gebel G, Maier J (2009) Poly(p-phenylene sulfone)s with high ion exchange capacity: ionomers with unique microstructural and transport features. *Phys Chem Chem Phys* 11:3305–3312
30. FumaPem - High performance membranes for fuel cells, Products section of company website [www.fumatech.com](http://www.fumatech.com)
31. Herz HG, Kreuer KD, Maier J, Scharfenberger G, Schuster MFH, Meyer WH (2003) New fully polymeric proton solvents with high proton mobility. *Electrochim Acta* 48:2165–2171
32. Scharfenberger G, Meyer WH, Wegner G, Schuster M, Kreuer KD, Maier J (2006) Anhydrous polymeric proton conductors based on imidazole functionalized polysiloxane. *Fuel Cells* 6:237–250
33. Schuster MFH, Meyer WH, Schuster M, Kreuer KD (2004) Toward a new type of anhydrous organic proton conductor based on immobilized imidazole. *Chem Mater* 16:329–337
34. Bozkurt A, Karadedeli B (2007) Copolymers of 4(5)-vinylimidazole and ethyleneglycol methacrylate phosphate: synthesis and proton conductivity properties. *React Funct Polym* 67:348–354
35. Steininger H, Schuster M, Kreuer KD, Maier J (2006) Intermediate temperature proton conductors based on phosphonic acid functionalized oligosiloxanes. *Solid State Ionics* 177:2457–2462
36. Bozkurt A, Meyer WH, Gutmann J, Wegner G (2003) Proton conducting copolymers on the basis of vinylphosphonic acid and 4-vinylimidazole. *Solid State Ionics* 164:169–176
37. Seel DC, Benicewicz BC, Xiao L, Schmidt TJ (2009) High-temperature polybenzimidazole-based membranes. In: Vielstich W, Yokokawa H, Gasteiger HA (eds) Handbook of Fuel Cells-Fundamentals, Technology and Applications, vol 5. John Wiley & Sons, Chichester, pp 300–312
38. PBI/H<sub>3</sub>PO<sub>4</sub> fuel cell starts up at room temperature. *Fuel Cells Bulletin* November 2008, p 10
39. Li Q, Jensen JO, Savinell RF, Bjerrum NJ (2009) High temperature proton exchange membranes based on polybenzimidazoles for fuel cells. *Prog Polym Sci* 34:449–477
40. Ahluwalia RK, Wang X (2006) Rapid self-start of polymer electrolyte fuel cell stacks from subfreezing temperatures. *J Power Sources* 162:502–512
41. Oszcipok M, Hakenjos A, Riemann D, Hebling C (2007) Start up and freezing processes in PEM fuel cells. *Fuel Cells* 7:135–141
42. Gebert M, Hoehlein B, Stolten D (2004) Benchmark cost analysis of main PEFC ionomer membrane solutions. *J Fuel Cell Sci Technol* 1:56
43. Springer TE, Wilson MS, Gottesfeld S (1993) Modeling and experimental diagnostics in polymer electrolyte fuel cells. *J Electrochem Soc* 140:3513–3526
44. Mathias MF, Roth J, Fleming J, Lehnert W (2003) Diffusion media materials and characterisation. In: Vielstich W, Gasteiger HA, Lamm A (eds) Handbook of fuel cells – fundamentals, technology and applications, Vol. 3. John Wiley & Sons, Chichester, pp 515–537
45. Neyerlin KC, Gu W, Jorne J, Gasteiger HA (2007) Study of the exchange current density for the hydrogen oxidation and evolution reactions. *J Electrochem Soc* 154:B631–B635
46. Gasteiger HA, Kocha SS, Sompalli B, Wagner FT (2005) Activity benchmarks and requirements for Pt, Pt-alloy, and non-Pt oxygen reduction catalysts for PEMFCs. *Appl Catal, B* 56:9–35
47. Markovic NM, Ross PN (2002) Surface science studies of model fuel cell electrocatalysts. *Surf Sci Rep* 45:117–229
48. Bonakdarpour A, Stevens K, Vernstrom GD, Atanasoski R, Schmoekkel AK, Debe MK, Dahn JR (2007) Oxygen reduction activity of Pt and Pt-Mn-Co electrocatalysts sputtered on nanostructured thin film support. *Electrochim Acta* 53:688–694
49. Debe MK, Schmoekkel AK, Vernstrom GD, Atanasoski R (2006) High voltage stability of nanostructured thin film catalysts for PEM fuel cells. *J Power Sources* 161:1002–1011
50. Debe MK (2003) Novel catalysts, catalyst support and catalyst coated membrane methods. In: Vielstich W, Gasteiger HA, Lamm A (eds) Handbook of fuel cells-fundamentals, technology and applications, vol 3. John Wiley & Sons, Chichester, pp 576–590
51. Gancs L, Kobayashi T, Debe MK, Atanasoski R, Wieckowski A (2008) Crystallographic characteristics of nanostructured thin-film fuel cell electrocatalysts: a HRTEM study. *Chem Mater* 20:2444–2454
52. Zhang JL, Vukmirovic MB, Xu Y, Mavrikakis M, Adzic RR (2005) Controlling the catalytic activity of platinum-monolayer electrocatalysts for oxygen reduction with different substrates. *Angew Chem Int Ed* 44:2132–2135
53. Stamenkovic VR, Mun BS, Mayrhofer KJJ, Ross PN, Markovic NM, Rossmeisl J, Greeley J, Norskov JK (2006) Changing the activity of electrocatalysts for oxygen reduction by tuning the surface electronic structure. *Angew Chem Int Ed* 45:2897–2901

54. Stamenkovic VR, Fowler B, Mun BS, Wang G, Ross PN, Lucas CA, Markovic NM (2007) Improved oxygen reduction activity on Pt<sub>3</sub>Ni(111) via increased surface site availability. *Science* 315:493–497
55. Antolini E, Salgado JRC, Gonzalez ER (2006) The stability of Pt-M (M=first row transition metal) alloy catalysts and its effect on the activity in low temperature fuel cells. *J Power Sources* 160:957–968
56. Mukerjee S, Srinivasan S (1993) Enhanced electrocatalysis of oxygen reduction on platinum alloys in proton exchange membrane fuel cells. *J Electroanal Chem* 357:201–224
57. Murthi VS (2009) Highly dispersed alloy catalyst for durability. DOE hydrogen programme FY 2009 annual progress report, pp 1075–1080
58. Adzic RR, Zhang J, Sasaki K, Vukmirovic MB, Shao M, Wang JX, Nilekar AU, Mavrikakis M, Valerio JA, Uribe F (2007) Platinum monolayer fuel cell electrocatalysts. *Top Catal* 46:249–262
59. Ball SC, Burton SL, Fisher J, ÓMalley R, Tessier BC, Theobald B, Thompsett D, Zhou WP, Su D, Zhu Y, Adzic R (2009) Structure and activity of novel Pt core-shell catalysts for the oxygen reduction reaction. *ECS Trans* 25:1023–1036
60. Neyerlin KC, Srivastava R, Yu C, Strasser P (2009) Electrochemical activity and stability of dealloyed Pt-Cu and Pt-Cu-Co electrocatalysts for the oxygen reduction reaction (ORR). *J Power Sources* 186:261–267
61. Strasser P (2009) Dealloyed Pt bimetallic electrocatalysts for oxygen reduction. In: Vielstich W, Yokokawa H, Gasteiger HA (eds) *Handbook of fuel cells-fundamentals, technology and applications*, vol 5. John Wiley & Sons, Chichester, pp 30–47
62. Wang X, Kariuki N, Niyogi S, Smith MC, Myers DJ, Hofmann T, Zhang Y, Bar M, Heske C (2008) Bimetallic palladium-base metal nanoparticle oxygen reduction electrocatalysts. *ECS Trans* 16:109–119
63. Zhou Y, Holme T, Berry J, Ohno TR, Ginley D, ÓHayre R (2009) Dopant-induced electronic structure modification of HOPG surfaces: implications for high activity fuel cell catalysts. *J Phys Chem C* 114:506–515
64. Shao Y, Liu J, Wang Y, Lin Y (2009) Novel catalyst support materials for PEM fuel cells: current status and future prospects. *J Mater Chem* 19:46–59
65. Bashyam R, Zelenay P (2006) A class of non-precious metal composite catalysts for fuel cells. *Nature* 443:63–66
66. Lefevre M, Proietti E, Jaouen F, Dodelet JP (2009) Iron-based catalysts with improved oxygen reduction activity in polymer electrolyte fuel cells. *Science* 324:71–74, Washington
67. Wu G, Artyushkova K, Ferrandon M, Kropf AJ, Myers D, Zelenay P (2009) Performance durability of polyaniline-derived non-precious cathode catalysts. *ECS Trans* 25:1299–1311
68. Neyerlin KC, Gasteiger HA, Mittelsteadt CK, Jorne J, Gu W (2005) Effect of relative humidity on oxygen reduction kinetics in a PEMFC. *J Electrochem Soc* 152:A1073–A1080
69. Kocha SS (2003) Principles of MEA preparation. In: Vielstich W, Gasteiger HA, Lamm A (eds) *Handbook of fuel cells-fundamentals, technology and applications*, vol 3. John Wiley and Sons, Chichester, UK, pp 538–565
70. Xie Z, Zhao X, Gazzarri J, Wang Q, Navessin T, Holdcroft S (2009) Identification of dominant transport mechanisms in PEMFC cathode catalyst layers operated under low RH. *ECS Trans* 25:1187–1192
71. Quick C, Ritzinger D, Lehnert W, Hartnig C (2009) Characterization of water transport in gas diffusion media. *J Power Sources* 190:110–120
72. Hermann A, Chaudhuri T, Spagnol P (2005) Bipolar plates for PEM fuel cells: a review. *Int J Hydrogen Energy* 30:1297–1302
73. Morikawa H, Kikushi H, Saito N (2009) Development and advances of a V-flow FC stack for FCX clarity. *SAE Int J Engines* 2:955–959
74. Shimoi R, Aoyama T, Iiyama A (2009) Development of fuel cell stack durability based on actual vehicle test data: current status and future work. *SAE Int J Engines* 2:960–970
75. Makkus RC, Janssen AHH, de Bruijn FA, Mallant RKAM (2000) Use of stainless steel for cost competitive bipolar plates in the SPFC. *J Power Sources* 86:274–282
76. Suria OV, Bruno M, Bois P, Maggiore P, Cazzolato C (2009) Fuel size and weight reduction due to innovative metallic bipolar plates: Technical process details and improvements, SAE Technical Papers Series, pp 2009-01-1009
77. Brady MP, Yang B, Wang H, Turner JA, More KL, Wilson M, Garzon F (2006) The formation of protective nitride surfaces for PEM fuel cell metallic bipolar plates. *JOM* 58:50–57
78. Cho EA, Jeon US, Hong SA, Oh IH, Kang SG (2005) Performance of a 1-kW-class PEMFC stack using TiN-coated 316 stainless steel bipolar plates. *J Power Sources* 142:177–183
79. Mepsted GO, Moore JM (2003) Performance and durability of bipolar plate materials. In: Vielstich W, Gasteiger HA, Lamm A (eds) *Handbook of fuel cells-fundamentals, technology and applications*, vol 3. John Wiley & Sons, Chichester, pp 286–293
80. Joseph S, McClure JC, Sebastian PJ, Moreira J, Valenzuela E (2008) Polyaniline and polypyrrole coatings on aluminum for PEM fuel cell bipolar plates. *J Power Sources* 177: 161–166
81. Ahluwalia R, Wang X, Lasher S, Sinha J, Yang Y, Sriramulu S (2007) Performance of automotive fuel cell systems with nanostructured thin film catalysts, 2007 fuel cell seminar, 15-10-2007. Courtesy Associates, Washington
82. Dobrovolskii YuA, Ukshe AE, Levchenko AE, Arkhangel'skii IV, Ionov SG, Avdeev VV, Aldoshin SM (2007) Materials for bipolar plates for proton-conducting membrane fuel cells. *Russ J Gen Chem* 77:752–765
83. Cleghorn SJC, Mayfield DK, Moore DA, Moore JC, Rusch G, Sherman TW, Sisofo N, Beuscher U (2006) A polymer electrolyte fuel cell life test: 3 years of continuous operation. *J Power Sources* 158:4–455
84. Schulze M, Knöri T, Schneider A, Gültow E (2004) Degradation of sealings for PEFC test cells during fuel cell operation. *J Power Sources* 127:222–229
85. Tan J, Chao YJ, Van Zee JW, Lee WK (2007) Degradation of elastomeric gasket materials in PEM fuel cells. *Mater Sci Eng A* 445–446:669–675

86. Ralph TR, Barnwell DE, Bouwman PJ, Hodgkinson AJ, Petch MI, Pollington M (2008) Reinforced membrane durability in proton exchange membrane fuel cell stacks for automotive applications. *J Electrochem Soc* 155: B411–B422
87. de Bruijn FA, Dam VAT, Janssen GJM (2008) Review: durability and degradation issues of PEM fuel cell components. *Fuel Cells* 8:3–22
88. Noto H, Kondo M, Otake Y, Kato M (2009) Development of fuel cell hybrid vehicle by Toyota, SAE technical paper series, pp 2009-01-1002
89. Reiser CA, Bregoli L, Patterson TW, Yi JS, Yang JD, Perry ML, Jarvi TD (2005) A reverse-current decay mechanism for fuel cells. *Electrochem Solid-State Lett* 8:A273–A276
90. Knights SD, Colbow KM, St-Pierre J, Wilkinson DP (2004) Aging mechanisms and lifetime of PEFC and DMFC. *J Power Sources* 127:127–134
91. Ferreira-Aparicio P, Chaparro AM, Gallardo B, Folgado M, Daza L (2009) Anode degradation effects in PEMFC stacks by localized fuel starvation, 2009 fuel cell seminar. Courtesy Associates, Washington
92. Schmittinger W, Vahidi A (2008) A review of the main parameters influencing long-term performance and durability of PEM fuel cells. *J Power Sources* 180:1–14
93. Baldasano JM, Valera E, Jimenez P (2003) Air quality data from large cities. *Sci Total Environ* 307:141–165
94. Huang W, Tan J, Kan H, Zhao N, Song W, Song G, Chen G, Jiang L, Jiang C, Chen R, Chen B (2009) Visibility, air quality and daily mortality in Shanghai, China. *Sci Total Environ* 407:3295–3300
95. A Report on the achievements and learnings from the HyFleet: CUTE Project 2006 – 2009
96. Narusawa K, Myong K, Murooka K, Kamiya Y (2007) A study regarding effects of proton exchange membrane fuel cell poisoning due to impurities on fuel cell performance, SAE technical paper series, pp 2007-01-0698
97. Adjemian K, Iiyama A (2008) MEA development for automotive applications, fuel cells durability and performance, 3rd edn. The Knowledge Press, Inc., Brookline, pp 5–16
98. Veldhuis JBJ, de Bruijn FA, Mallant RKAM (1998) Fuel cell seminar abstracts, 16-11-1998 Courtesy Associates, Washington, pp 598
99. Knights SD, Jia N, Chuy C, Zhang J (2005) Fuel cell seminar abstracts, 14-11-2005 Courtesy associates, Washington
100. Cheng X, Shi Z, Glass N, Zhang L, Zhang J, Song D, Liu ZS, Wang H, Shen J (2007) A review of PEM hydrogen fuel cell contamination: Impacts, mechanisms, and mitigation. *J Power Sources* 165:739–756
101. Kennedy DM, Cahela DR, Zhu WH, Westrom KC, Nelms RM, Tatarchuk BJ (2007) Fuel cell cathode air filters: methodologies for design and optimization. *J Power Sources* 168:391–399
102. Matsuda Y (2009) Accumulation behavior of impurities in fuel cell hydrogen circulation system, 2009 fuel cell seminar, 16-11-2009 Courtesy associates, Washington
103. Papasavva S (2005) Developing hydrogen (H<sub>2</sub>) specification guidelines for proton exchange membrane (PEM) fuel cell vehicles, SAE technical series papers, pp 2005-01-0011
104. Recupero V, Pino L, Vita A, Cipiti F, Cordaro M, Lagana M (2005) Development of a LPG fuel processor for PEFC systems: Laboratory scale evaluation of autothermal reforming and preferential oxidation subunits. *Int J Hydrogen Energy* 30:963–971
105. de Bruijn FA, Rietveld G, van den Brink RW (2007) Hydrogen production and fuel cells as the bridging technologies towards a sustainable energy system. In: Centi G, Santen RA (eds) *Catalysis for renewables*. Wiley-VCH, Weinheim, pp 299–336
106. Li Q, He R, Gao J-A, Jensen JO, Bjerrum NJ (2003) The CO poisoning effect in polymer electrolyte fuel cells operational at temperatures up to 200°C. *J Electrochem Soc* 150:A1599–A1605
107. Wipke K, Sprick S, Kurtz J, Ramsden T (2009) Controlled hydrogen fleet and infrastructure demonstration and validation project, NREL/NREL/TP-560-46679
108. Mallant RKAM, Lebedeva NP, Zhang YM, Li L, Tang JK, Bukhtiyarov VI, Romanenko AV, Voropaev I, Bauer B, Zhang W, Jones DJ, Rozière J, Gao HR (2009) Significant steps towards medium temperature/low RH PEMFC, 2009 fuel cell seminar. Courtesy Associates, Washington
109. Bono T, Kizaki M, Mizuno H, Nonobe Y, Takahashi T, Matsumoto T, Kobayashi N (2010) Development of new Toyota FCHV-adv fuel cell system. *SAE Int J Engines* 2:948–954
110. Power backup solutions for telecom and related networks. Dantherm Power Catalogue 2008
111. Ferreira PJ, la Ó GJ, Shao-Horn Y, Morgan D, Makharia R, Kocha S, Gasteiger HA (2005) Instability of Pt/C electrocatalysts in proton exchange membrane fuel cells. *J Electrochem Soc* 152:A2256–A2271
112. Xie J, Wood DL, More KL, Atanassov P, Borup RL (2005) Microstructural changes of membrane electrode assemblies during PEFC durability testing at high humidity conditions. *J Electrochem Soc* 152:A1011–A1020
113. Guilminot E, Corcella A, Charlot F, Maillard F, Chatenet M (2007) Detection of Pt[sup z+] ions and Pt nanoparticles inside the membrane of a used PEMFC. *J Electrochem Soc* 154:B96–B105
114. Shao-Horn Y, Sheng WC, Chen S, Ferreira PJ, Holby EF, Morgan D (2007) Instability of supported platinum nanoparticles in low-temperature fuel cells. *Top Catal* 46:285–305
115. Pourbaix M (1974) Atlas of electrochemical equilibria in aqueous solutions. National Association of Corrosion Engineers, New York
116. Darling RM, Meyers JP (2003) Kinetic model of platinum dissolution in PEMFCs. *J Electrochem Soc* 150:A1523–A1527
117. Virkar AN, Zhou Y (2007) Mechanism of catalyst degradation in proton exchange membrane fuel cells. *J Electrochem Soc* 154:B540–B547
118. Dam VAT, de Bruijn FA (2007) The stability of PEMFC electrodes. *J Electrochem Soc* 154:B494–B499

119. Wang X, Kumar R, Myers DJ (2006) Effect of voltage on platinum dissolution. *Electrochem Solid-State Lett* 9:A225–A227
120. Kawahara S, Mitsushima S, Ota K, Kamiya N (2006) Deterioration of Pt catalyst under potential cycling. *ECS Trans* 3:625–631
121. Kawahara S, Mitsushima S, Ota K, Kamiya N (2006) Consumption of Pt catalyst under electrolysis and fuel cell operation. *ECS Trans* 1:85–100
122. Burke LD, Buckley DT (1994) Anomalous stability of acid-grown hydrous platinum oxide films in aqueous media. *J Electroanal Chem* 366:239–251
123. Burke LD, ÓDwyer KJ (1992) Multilayer oxide growth on Pt under potential cycling conditions. *Electrochim Acta* 37:43–50
124. Birss VI, Chang M, Segal J (1993) Platinum oxide film formation-reduction: an in-situ mass measurement study. *J Electroanal Chem* 355:181–191
125. Nagy Z, You H (2002) Applications of surface X-ray scattering to electrochemistry problems. *Electrochim Acta* 47:3037–3055
126. Kinoshita K (1992) *Electrochemical oxygen technology*. John Wiley & Sons, Inc., New York
127. Guilminot E, Corcella A, Chatenet M, Maillard F, Charlot F, Berthome G, Iojoiu C, Sanchez JY, Rossinot E, Claude E (2007) Membrane and active layer degradation upon PEMFC steady-state operation. *J Electrochem Soc* 154:B1106–B1114
128. Honji A, Mori T, Tamura K, Hishinuma M (1988) Agglomeration of platinum particles supported on carbon in phosphoric acid. *J Electrochem Soc* 135:355–359
129. Ascarelli P, Contini V, Giorgi R (2002) Formation process of nanocrystalline materials from x-ray diffraction profile analysis: application to platinum catalysts. *J Appl Phys* 91:4556–4561
130. Borup RL, Davey JR, Garzon FH, Wood DL, Inbody MA (2006) PEM fuel cell electrocatalyst durability measurements. *J Power Sources* 163:76–81
131. Mathias MF, Makharia R, Gasteiger HA, Conley JJ, Fuller TJ, Gittleman CJ, Kocha SS, Miller DP, Mittelsteadt CK, Xie T, Yan SG, Yu PT (2005) Two fuel cell cars in every garage. *Electrochem Soc Interface* 14(Fall):24–35
132. Haas HR, Davis MT (2009) Electrode and catalyst durability requirements in automotive PEM applications: technology status of a recent MEA design and next generation challenges. *ECS Trans* 25:1623–1631
133. Schulze M, Wagner N, Kaz T, Friedrich KA (2007) Combined electrochemical and surface analysis investigation of degradation processes in polymer electrolyte membrane fuel cells. *Electrochim Acta* 52:2328–2336
134. Janssen GJM, de Heer MP, Papageorgopoulos DC (2004) Bilayer anodes for improved reformate tolerance of PEM fuel cells. *Fuel Cells* 4:169–174
135. Yu H, Hou Z, Yi B, Lin Z (2002) Composite anode for CO tolerance proton exchange membrane fuel cells. *J Power Sources* 105:52–57
136. Pielak P, Eickes C, Brosha E, Garzon F, Zelenay P (2004) Ruthenium crossover in direct methanol fuel cell with Pt-Ru black anode. *J Electrochem Soc* 151:A2053–A2059
137. Lebedeva NP, Rosca V, Janssen GJM (2010) CO oxidation and CO<sub>2</sub> reduction on carbon supported PtWO<sub>3</sub> catalyst. *Electrochim Acta* 55:7659–7668
138. de Bruijn FA, Papageorgopoulos DC, Sitters EF, Janssen GJM (2002) The influence of carbon dioxide on PEM fuel cells anodes. *J Power Sources* 110:117–124
139. Janssen GJM (2004) Modelling study of CO<sub>2</sub> poisoning on PEMFC anodes. *J Power Sources* 136:45–54
140. Ahluwalia RK, Wang X (2008) Effect of CO and CO<sub>2</sub> impurities on performance of direct hydrogen polymer-electrolyte fuel cells. *J Power Sources* 180:122–131
141. Mohtadi R, Lee W, Van Zee JW (2004) Assessing durability of cathodes exposed to common air impurities. *J Power Sources* 138:216–225
142. Paulus UA, Schmidt TJ, Gasteiger HA (2003) Poisons for the O<sub>2</sub> reduction reaction. In: Vielstich W, Gasteiger HA, Lamm A (eds) *Handbook of fuel cells-fundamentals, technology and applications*, vol 2. John Wiley & Sons, Chichester, pp 555–569
143. Kinoshita K (1988) *Carbon. Electrochemical and physicochemical properties*. John Wiley & Sons, Inc., New York
144. Giordano N, Antonucci PL, Passalacqua E, Pino L, Arico AS, Kinoshita K (1991) Relationship between physicochemical properties and electrooxidation behaviour of carbon materials. *Electrochim Acta* 36:1931–1935
145. Ball SC, Hudson SL, Thompsett D, Theobald B (2007) An investigation into factors affecting the stability of carbons and carbon supported platinum and platinum/cobalt alloy catalysts during 1.2 V potentiostatic hold regimes at a range of temperatures. *J Power Sources* 171:18–15
146. Garland N, Benjamin T, Kopasz J (2007) DOE fuel cell program: durability technical targets and testing protocols. *ECS Trans* 11:923–931
147. Stevens DA, Dahn JR (2005) Thermal degradation of the support in carbon-supported platinum electrocatalysts for PEM fuel cells. *Carbon* 43:179–188
148. Stevens DA, Hicks MT, Haugen GM, Dahn JR (2005) Ex situ and in situ stability studies of PEMFC catalysts. *J Electrochem Soc* 152:A2309–A2315
149. Cai M, Ruthkosky MS, Merzougui B, Swathirajan S, Balogh MP, Oh SE (2006) Investigation of thermal and electrochemical degradation of fuel cell catalysts. *J Power Sources* 160:977–986
150. Shao Y, Yin G, Gao Y, Shi P (2006) Durability study of Pt/C and Pt/CNTs catalysts under simulated PEM fuel cell conditions. *J Electrochem Soc* 153:A1093–A1097
151. Tang Z, Ng HY, Lin J, Wee ATS, Chua DHC (2010) Pt/CNT-based electrodes with high electrochemical activity and stability for proton exchange membrane fuel cells. *J Electrochem Soc* 157:B245–B250
152. Healy J, Hayden C, Xie T, Olson K, Waldo R, Brundage M, Gasteiger H, Abbott J (2005) Aspects of the chemical degradation of PFSA ionomers used in PEM fuel cells. *Fuel Cells* 5:302–308
153. St-Pierre J, Wilkinson DP, Knights SD, Bos M (2000) Relationships between water management, contamination and

- lifetime degradation in PEFC. *J New Mater Electrochem Syst* 3:99–106
154. Wood D, Davey J, Garzon F, Atanassov P, Borup R (2005) Mass-transport phenomena and long-term performance limitations in H<sub>2</sub>-air PEMFC durability testing, 2005 fuel cell seminar abstracts, 14-11-2005 Courtesy Associates, Washington
  155. Jordan LR, Shukla AK, Behrsing T, Avery NR, Muddle BC, Forsyth M (2000) Diffusion layer parameters influencing optimal fuel cell performance. *J Power Sources* 86:250–254
  156. Williams MV, Begg E, Bonville L, Kunz HR, Fenton JM (2004) Characterization of gas diffusion layers for PEMFC. *J Electrochem Soc* 151:A1173–A1180
  157. de Bruijn FA, Dam VAT, Janssen GJM, Makkus RC (2009) Electrode degradation in PEMFCs as studied in model systems and PEMFC testing. *ECS Trans* 25:1835–1847
  158. Lee C, Merida W (2007) Gas diffusion layer durability under steady-state and freezing conditions. *J Power Sources* 164:141–153
  159. Coms FD (2008) The chemistry of fuel cell membrane chemical degradation. *ECS Trans* 16:235–255
  160. Liu H, Gasteiger HA, LaConti AB, Zhang J (2006) Factors impacting chemical degradation of perfluorinated sulfonic acid ionomers. *ECS Trans* 1:283–293
  161. Mittal VO, Kunz HR, Fenton JM (2007) Membrane degradation mechanisms in PEMFCs. *J Electrochem Soc* 154:B652–B656
  162. Curtin DE, Lousenberg RD, Henry TJ, Tangeman PC, Tisack ME (2004) Advanced materials for improved PEMFC performance and life. *J Power Sources* 131:41–48
  163. Coms FD, Liu H, Owejan JE (2008) Mitigation of perfluoro-sulfonic acid membrane chemical degradation using cerium and manganese ions. *ECS Trans* 16:1735–1747
  164. Endoh E (2008) Development of highly durable PFSA membrane and MEA for PEMFC under high temperature and low humidity conditions. *ECS Trans* 16:1229–1240
  165. Trogadas P, Parrondo J, Ramani V (2008) Degradation mitigation in polymer electrolyte membranes using free radical scavengers. *ECS Trans* 16:1725–1733
  166. Rozière J, Jones DJ (2003) Non-fluorinated polymer materials for proton exchange membrane fuel cells. *Annu Rev Mater Res* 33:503–555
  167. Zhang L, Ma CS, Mukerjee S (2003) Oxygen permeation studies on alternative proton exchange membranes designed for elevated temperature operation. *Electrochim Acta* 48:1845–1859
  168. Schuster M, Kreuer KD, Andersen HT, Maier J (2007) Sulfonated poly(phenylene sulfone) polymers as hydrolytically and thermooxidatively stable proton conducting ionomers. *Macromolecules* 40:598–607
  169. Escobedo G, Raiford K, Nagarajan GS, Schwiebert KE (2006) Strategies for mitigation of PFSA polymer degradation in PEM fuel cells. *ECS Trans* 1:303–311
  170. Stone C, Calis GHM (2006) Improved composite membranes and related performance in commercial PEM fuel cells, 2006 fuel cell seminar abstracts. Courtesy Associates, Washington
  171. LaConti AB, Hamdan M, McDonald RC (2003) Mechanisms of membrane degradation. In: Vielstich W, Lamm A, Gasteiger HA (eds) *Handbook of fuel cells*, 3. Wiley, Chichester, pp 647–663
  172. Silberstein MN, Boyce MC (2010) Constitutive modeling of the rate, temperature, and hydration dependent deformation response of Nafion to monotonic and cyclic loading. *J Power Sources* 195:5692–5706
  173. Liu W, Ruth K, Rusch G (2001) Membrane durability in PEM fuel cells. *J New Mat Electrochem Syst* 4:227–231
  174. McDonald RC, Mittelsteadt CK, Thompson EL (2004) Effects of deep temperature cycling on Nafion® 112 membranes and membrane electrode assemblies. *Fuel Cells* 4:208–213
  175. Okada T (2003) Ionic Contaminants. In: Vielstich W, Lamm A, Gasteiger HA (eds) *Handbook of fuel cells*, 3. Wiley, Chichester, pp 627–646
  176. Davies DP, Adcock PL, Turpin M, Rowen SJ (2000) Bipolar plate materials for solid polymer fuel cells. *J Appl Electrochem* 30:101–105
  177. Gallagher KG, Wong DT, Fuller TF (2008) The effect of transient potential exposure on the electrochemical oxidation of carbon black in low-temperature fuel cells. *J Electrochem Soc* 155:B488–B493
  178. Frisch L (2001) PEM fuel cell stack sealing using silicone elastomers. *Sealing Technol* 2001:7–9
  179. Du B, Guo R, Pollard R, Rodriguez D, Smith J, Elter J (2006) PEM fuel cells: status and challenges for commercial stationary power applications. *JOM* 8:45–49
  180. St-Pierre J, Jia N (2002) Successful demonstration of Ballard PEMFCs for space shuttle applications. *J New Mater Electrochem Syst* 5:263

## Books and Reviews

- Barbir F (2005) *PEM fuel cells, theory and practice*. Elsevier, Amsterdam
- Büchi FN, Inaba M, Schmidt TJ (eds) (2009) *Polymer electrolyte fuel cell durability*. Springer, New York
- Larminie J, Dicks A (2003) *Fuel cell systems explained*, 2nd edn. Wiley, Chichester
- Scherer GG (ed) (2008) *Fuel cells I, advances in polymer science*, vol 215. Springer, New York
- Scherer GG (ed) (2008) *Fuel cells II, advances in polymer science*, vol 216. Springer, New York
- Vielstich W, Lamm A, Gasteiger HA (eds) (2003) *Handbook of fuel cells, fundamental, technology and applications*, 4th edn. Wiley, Chichester
- Vielstich W, Yokokawa H, Gasteiger HA (eds) (2009) *Handbook of fuel cells: advances in electrocatalysis, materials, diagnostics and durability*, vol 5 & 6. Wiley, Chichester
- Zhang J (ed) (2009) *PEM fuel cell electrocatalysts and catalyst layers*. Springer, New York
- Zhao TS, Kreuer KD, Van Nguyen T (eds) (2007) *Advances in fuel cells I*. Elsevier, Amsterdam



## PEM Fuel Cells and Platinum-Based Electrocatalysts

JUNLIANG ZHANG

Electrochemical Energy Research Laboratory, General Motors Global R&D, Honeoye Falls, NY, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Electrocatalysis of the ORR at Platinum Surfaces

Pt-Alloy Electrocatalysts

Pt Monolayer Electrocatalysts

Pt and Pt-Alloy Nanowire and Nanotube

Electrocatalysts

Facet- and Shape-Controlled Pt-Alloy Nanocrystal

Electrocatalysts

Future Directions

Bibliography

### Glossary

**Anode** An electrode where the electrochemical oxidation reaction(s) occurs, generating free electrons that flow through a polarized electrical device and enter the cathode. In a fuel cell, the fuel oxidation reaction happens at the anode.

**Cathode** An electrode where the electrochemical reduction reaction(s) occurs, by consuming the electrons originated from the anode. In a fuel cell, the oxygen reduction reaction happens at the cathode.

**Electrocatalyst** A material that is applied on the surface of an electrode to catalyze half-cell reactions.

**Normal hydrogen electrode (NHE)** Also known as the standard hydrogen electrode (SHE), it is a redox reference electrode which forms the basis of the thermodynamic scale of oxidation–reduction potentials. The potential of the NHE is defined as zero and based on equilibrium of the following redox half-cell reaction, typically on a Pt surface:  $2\text{H}^+(\text{aq}) + 2\text{e}^- \rightarrow \text{H}_2(\text{g})$  The activities of both the reduced form and the oxidized form are maintained at unity. That implies that the pressure

of hydrogen gas is 1 atm and the concentration of hydrogen ions in the solution is 1 M.

**Oxygen reduction reaction (ORR)** An electrode reaction, in which oxygen gas is reduced at the cathode of an electrochemical cell. The product of the reaction can be water molecules, hydroxyl ions ( $\text{OH}^-$ ), or sometimes hydrogen peroxide molecules. It is a very important and much-studied electrochemical reaction because it occurs at the cathode of practically all fuel cells.

**Proton-exchange membrane fuel cells (PEMFC)**

Also known as polymer electrolyte membrane fuel cells, these are a type of fuel cells that use proton-conducting-ionomer membrane as the electrolyte to separate anode and cathode. Their distinguishing features include low operating temperature ( $\sim 80^\circ\text{C}$ ), high power density, quick start-up, and quick match to shifting demands for power. They are being developed for transport applications as well as stationary and portable applications.

**Pt mass activity** The kinetic current of the oxygen reduction reaction normalized by the mass of Pt metal contained in the electrocatalyst.

**Pt-specific activity** The kinetic current of the oxygen reduction reaction normalized by the electrochemical surface area of the Pt metal contained in the electrocatalyst.

**Reversible hydrogen electrode (RHE)** This differs from the NHE by the fact that the hydrogen-ion concentration of RHE reaction is the same as that in the actual electrolyte solution used for the working electrode. The potential of RHE is therefore  $(-0.059^* (\text{pH of the electrolyte})) \text{ V}$ .

### Definition of the Subject

$\text{H}_2/\text{air}$ -powered PEM fuel cells are a future substitute for combustion engines as the green power source for transport application. In PEM fuel cells, because of their low operating temperature and low pH, both anode and cathode reactions are catalyzed by Pt or Pt-based electrocatalysts. Pt is a precious and expensive noble metal, and therefore its loading requirement plays a major role in determining the cost of fuel cells in mass production. The anode hydrogen oxidation reaction on Pt is intrinsically fast and requires very little Pt, while the cathode oxygen reduction reaction (ORR)

is a very sluggish reaction that consumes about 90% of the total Pt content in PEM fuel cells. The current Pt loading in the most advanced fuel cell vehicles that use state-of-the-art Pt-based catalysts is about four- to eightfold higher than the target established for mass-produced fuel cell vehicles. Therefore, lowering the Pt loading at the cathode is the most critical mission for the PEM fuel cell development. To do that, significant depth of knowledge in understanding the ORR on Pt and Pt-based electrocatalysts' surfaces is required; and the search for novel Pt-based electrocatalysts with enhanced ORR activity is seemingly the most productive pathway.

## Introduction

Ever since 1839, when Sir William Robert Grove introduced the first concept of fuel cells [1, 2], researchers have been continuously trying to apply fuel cells for stationary and mobile power sources [3, 4], because of their high energy efficiency and low environmental footprint. Fuel cells can be customarily classified according to the electrolyte employed, with different electrolytes operating at different temperature ranges. The operating temperature then dictates the types of fuels and electrode materials that can be used in a fuel cell. For example, aqueous electrolytes are limited to operating temperatures of about 200°C or lower because of their high water vapor pressure. At these temperatures the fuel is, in applications requiring high current density and low cost, restricted to hydrogen. To accommodate the slow kinetics of the electrochemical reactions at such low temperature in acid environments, platinum catalysts are required for both cathode and anode. In high temperature fuel cells, CO or even CH<sub>4</sub> can be used as the fuel, and the catalyst is not necessary to be noble metals, because of the inherently fast electrode reaction kinetics.

Among various types of fuel cells, proton exchange membrane fuel cells (PEMFCs) have attracted the most attention in recent years due to their low operating temperature (about 80°C), high power density, quick start-up, and quick match to shifting demands for power [3]. Hydrogen/air-powered PEMFCs make them the primary candidate for the power source of light-duty vehicles and buildings. In a recent study by Thomas [5], the author compared fuel cell and battery

as the power sources for all-electric vehicles, and found that for any vehicle range greater than 160 km (100 miles), fuel cells are superior than batteries in terms of mass, volume, cost, initial greenhouse gas reductions, refueling time, well-to-wheel energy efficiency (using natural gas and biomass as the source) and life cycle costs. PEMFCs also allow direct use of methanol without a processor, called DMFC. DMFCs are the primary candidates for portable electronic applications; low power densities and high Pt requirements have precluded their use in vehicles to date.

One difficult challenge to make PEMFC vehicles cost-competitive with traditional combustion engine cars is the high platinum catalyst loading and poor catalyst durability of the fuel cells [6, 7]. In the mid- to late-1980s and in the early 1990s, the fuel cell team at Los Alamos National Laboratory (LANL) succeeded in demonstrating high performance H<sub>2</sub>/air fuel cells with a platinum loading less than 0.5 mg Pt/cm<sup>2</sup> per electrode, which was one magnitude lower than its previous level [8–11]. In that effort, Raistrick [8] was the first one to cast ionomer into the electrocatalyst layer by impregnating a Pt/C electrode into the ionomer solution before hot-pressing it onto the membrane, and thus greatly increased the electrocatalyst–electrolyte interfacial contact area. Wilson et al. [9–11] improved that process by mixing the Pt/C powder catalyst with ionomer solution before coating the electrode. Those research achievements at LANL partly led to the renaissance in PEMFCs in the past 2 decades [4].

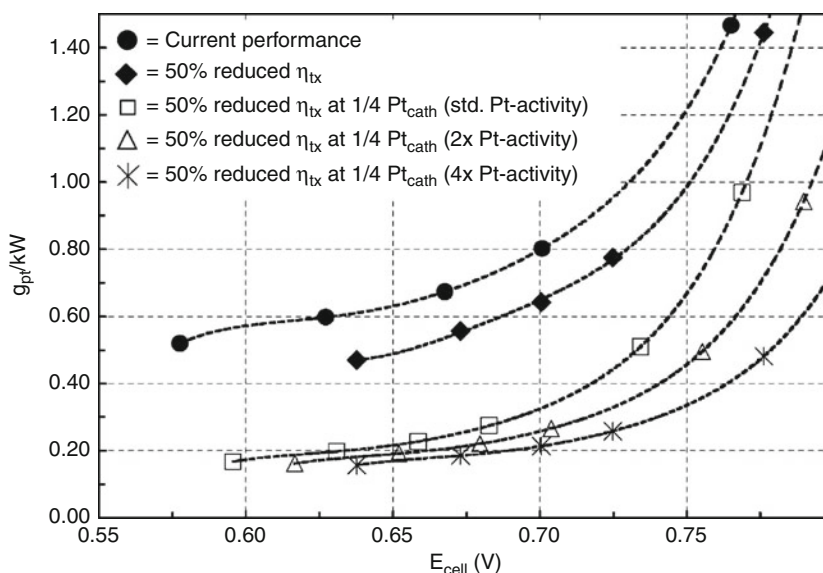
In a PEMFC, if pure hydrogen is used as fuel, the anode reaction is then the hydrogen oxidation reaction (HOR) at the surface of the anode platinum electrocatalyst. The hydrogen oxidation reaction (HOR) and hydrogen evolution reaction (HER) are by far the most thoroughly investigated electrochemical reaction system [12]. Due to the fast electrode kinetics of hydrogen oxidation at platinum surface [12–14], the anode platinum loading can be reduced down to 0.05 mg Pt/cm<sup>2</sup> without significant performance loss [15]. The cathode reaction in a PEM fuel cell is the oxygen reduction reaction (ORR) at platinum surface in an acidic electrolyte. In contrast to HOR at the anode, the cathode ORR is a highly irreversible reaction even at temperatures above 100°C at the best existing catalyst – the platinum surface [16–19]. Gasteiger et al. [20] found that 0.4 mg Pt/cm<sup>2</sup> was close to the optimal

platinum loading for the air electrode using the state-of-the-art Pt/C catalyst and an optimized electrode structure. Further reduction of the cathode platinum loading will result in cell voltage loss at low current densities that follows the ORR kinetic loss. The high platinum loading at the cathode originates from the slow kinetics of (ORR) at platinum surface. To make the fuel cell vehicles commercially viable on the market, the platinum loading on the cathode has to be reduced significantly. As shown in Fig. 1, when Pt loading requirement is translated from the target set for power-specific Pt consumption in g Pt/kW, a fourfold Pt mass activity (activity per unit platinum mass) improvement is required, if combined with a 50% reduction in mass transport-related voltage loss [21]. Recent increases in Pt prices suggest that one should be striving for at least an eightfold improvement.

There are two ways one might think of that could help to reach that goal: further increasing the platinum dispersion (defined as the ratio of surface metal atoms

to total number of atoms) by making finer platinum particles, if there is no decrease of Pt-specific activity (activity per unit Pt surface area); or alternatively, increasing the Pt-specific activity. One could also seek a combination of these two approaches.

This selected brief review will be focused on the research and development progress on ORR kinetics. The origin of the problem related with the low ORR activity of platinum will be discussed, followed by a review of recent progress in making more active, more durable platinum-based ORR catalysts. These include platinum alloy catalysts, platinum monolayer catalysts, platinum nanowire and nanotube catalysts, and the more recent shape- and facet-controlled platinum-alloy nanocrystal catalysts. The progress in the mechanistic understanding on the correlation between the activity and the electronic and structural properties of surface platinum atoms will be reviewed as well. The future direction of the research on platinum-based catalysts for PEM fuel cell application will be proposed.

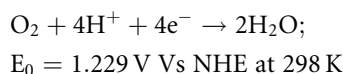


**PEM Fuel Cells and Platinum-Based Electrocatalysts. Figure 1**

Pt-mass-specific power density [ $g_{Pt}/kW$ ] versus cell voltage,  $E_{cell}$  [V], based on a  $50\text{ cm}^2$  single-cell  $H_2$ /air performance.  $\eta_{tx}$  = cell-voltage loss caused by mass transport;  $Pt_{cath}$  = cathode-Pt loading. The cell was tested at  $T_{cell} = 80^\circ\text{C}$ , 100% RH (relative humidity), at a total pressure of  $150\text{ kPa}_{abs}$  and stoichiometric flows of  $s = 2.0/2.0$ . Catalyst-coated membrane (CCM) was based on a ca.  $25\text{ }\mu\text{m}$  low-EW (equivalent weight = 900) membrane, and ca. 50 wt% Pt/carbon ( $0.4/0.4\text{ mg Pt}/\text{cm}^2$  (anode/cathode)). It was assumed that the cell performance could be maintained at a reduced anode loading of  $0.05\text{ mg Pt}/\text{cm}^2$  (Reproduced from [21]. With permission)

### Electrocatalysis of the ORR at Platinum Surfaces

It is widely accepted that the ORR on platinum surfaces is dominantly a multistep and four-electron reduction process with  $\text{H}_2\text{O}$  being the final product. However, the detailed mechanism of ORR still remains elusive [17]. The overall four electron reduction of  $\text{O}_2$  in acid aqueous solutions is



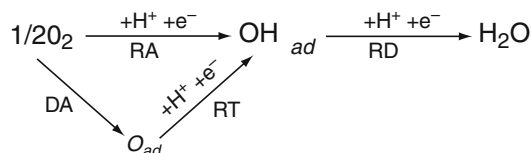
Since the four-electron reduction of oxygen is highly irreversible, the experimental verification of the thermodynamic reversible potential of this reaction is very difficult. The irreversibility of ORR imposes serious voltage loss in fuel cells. In most instances, the current densities practical for kinetic studies are much larger than the exchange current density of ORR; therefore the information obtained from current-potential data are confined only to the rate-determining step (RDS). On the other hand, in the ORR kinetic potential region, the electrode surface structure and properties strongly depend on the applied potential and the time held at that potential, which makes the reaction more complicated. While the relationship between the overall kinetics and the surface electronic properties is not well understood, it is widely accepted that in the multistep reaction, the first electron transfer is the rate-determining step, which is accompanied by or followed by a fast proton transfer [16–18]. Two Tafel slopes are usually observed for ORR on Pt in RDE tests in perchloric acid, from  $-60 \text{ mV/decade}$  at low current density, transitioning to  $-120 \text{ mV/decade}$  at high current density. The lower Tafel slope of the ORR in perchloric acid at low current density has been attributed to the potential dependent Pt oxide/hydroxide coverage at high potentials [22–26].

Recently, by using Density Functional Theory (DFT), Norskov and coworkers [27] calculated the Gibbs free energy of ORR intermediates as a function of cathode potential based on a simple dissociative mechanism, i.e., with the adsorbed oxygen and hydroxide being the only intermediates. They found that oxygen or hydroxide is so strongly bound to the platinum surface at the thermodynamic equilibrium potential that proton and electron transfer become impossible. By lowering the potential, the stability of adsorbed

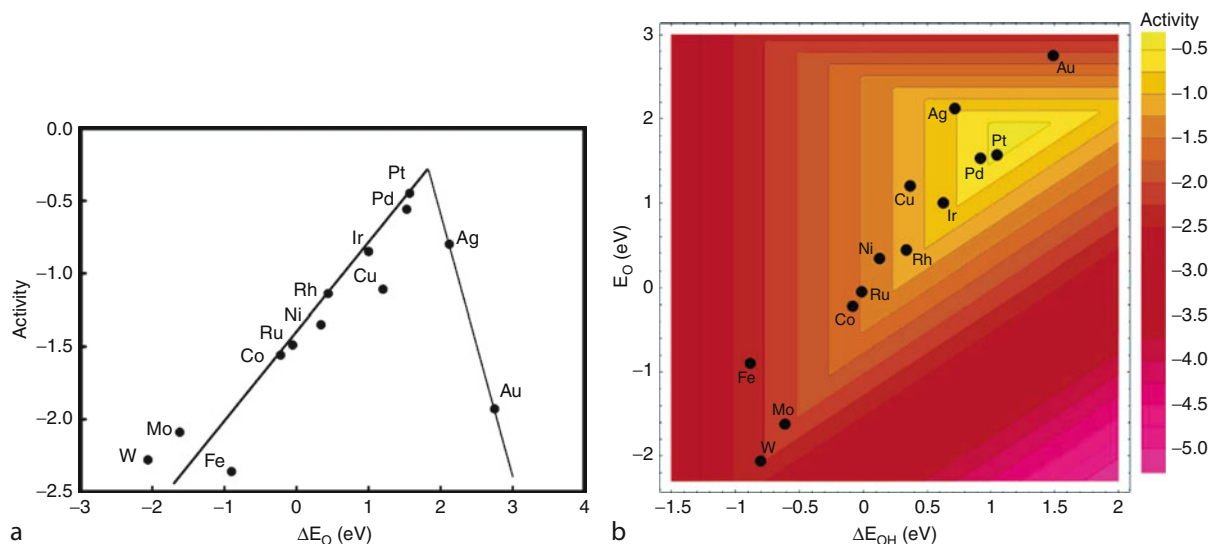
oxygen decreases and the reaction may proceed. They suggested that these effects are the origin of the overpotential of the ORR on platinum surfaces.

By setting the reference zero potential to be NHE, the proton chemical potential in the electrolyte is related to the electrode potential. The authors made DFT calculations to get the bond energies of  $\text{O}^*$  and  $\text{HO}^*$  for a number of interesting metals. From this, they can evaluate the reaction free energies of the two basic steps: the hydrogenation of the two adsorbed intermediates  $\text{O}^*$  and  $\text{HO}^*$ . The larger one of the two reaction free energies was taken as the activation energy barrier of the rate-determining step (RDS) in ORR. By using a microkinetic model they constructed, the rate constant of the RDS, and therefore the ORR activity, can be evaluated based on the activation energy barrier. As shown in Fig. 2a, the model predicts a volcano-shaped relationship between the rate of the ORR and the oxygen adsorption energy, with platinum and palladium being among the best metals for electrocatalysis of ORR. Figure 2b shows that the bonding energy of OH is roughly linearly correlated to that of O, indicating both are nearly equivalent parameters in determining the ORR activity.

More recently, Wang et al. [28] derived an intrinsic kinetic equation for the four-electron ( $4\text{e}^-$ ) oxygen reduction reaction (ORR) in acidic media, by using free energies of activation and adsorption as the kinetic parameters, which were obtained through fitting experimental ORR data from a Pt(111) rotating disk electrode (RDE). Their kinetic model consists of four essential elementary reactions: (1) a dissociative adsorption (DA); (2) a reductive adsorption (RA), which yields two reaction intermediates, O and OH; (3) a reductive transition (RT) from O to OH; and (4) a reductive desorption (RD) of OH, as shown below [28] (Reproduced with permission from [28]).



In contrast to a conventional ORR kinetic model, in this work there is no single particular RDS assumed, as the authors believe that a single RDS assumption may



PEM Fuel Cells and Platinum-Based Electrocatalysts. **Figure 2**

(a) Trends in oxygen reduction activity plotted as a function of the O-binding energy. (b) Trends in oxygen reduction activity plotted as a function of both the O and the OH-binding energy (Reproduced from [27]. With permission)

not hold over a wide potential region since the reaction pathway may change with potential, or there may exist two RDSs with similarly high activation barriers. The results indicate that the first electron transfer in the RA reaction ( $\Delta G_{\text{RA}}^{*0} = 0.46\text{eV}$ , where  $\Delta G_i^{*0}$  is the activation energy of reaction  $i$  (at equilibrium potential)) is not the rate-determining step (RDS) for the ORR on Pt at high potentials, because dissociative adsorption ( $\Delta G_{\text{DA}}^{*0} = 0.26\text{eV}$ ) provides a more active adsorption pathway. However, the reaction intermediates, O and OH, are strongly trapped on the Pt surface, requiring considerable overpotential to overcome the barriers for O to OH transition ( $\Delta G_{\text{RT}}^{*0} = 0.50\text{eV}$ ) and OH reduction to water and desorption ( $\Delta G_{\text{RD}}^{*0} = 0.45\text{eV}$ ). Thus, the ORR on Pt is desorption-limited at high potentials, exhibiting a low apparent Tafel slope at those potentials. Wang et al. [29] further used this kinetic model to fit a typical IR-free polarization curve of a PEMFC, by adjusting the parameters to reflect the fuel cell-operating conditions at  $80^\circ\text{C}$ . The results showed that the transition of the Tafel slope occurs at about the same 0.77 V that is the equilibrium potential for the transition between adsorbed O and OH on a Pt surface with low adsorbed O coverage [27].

Neyerlin et al. [30–32] investigated the ORR kinetics on high-surface-area carbon-supported platinum

catalyst Pt/C in an operating PEMFC. By assuming the transfer coefficient  $\alpha = 1$  and using a single Tafel slope, i.e.,  $-70\text{ mV/decade}$  at  $80^\circ\text{C}$ , three kinetic parameters could be extracted through fitting the kinetic model to the fuel cell data: the exchange current density or current density at a constant IR-free cell potential, the reaction order with respect to oxygen partial pressure, and the activation energy. One may need to note that the lower limit of the electrode potential after IR and transport correction in this work is about 0.77 V, so that this single-Tafel-slope treatment can still be consistent with Wang et al.'s [29] result. Neyerlin et al. [30] were concerned about the accuracy of the kinetic current extracted from low potentials in RDE tests, since at these low potentials the experimental measured current is more than ten times lower than the kinetic current derived from it. The transport correction used in RDE analysis assumes perfect first-order kinetics, which is not strictly true, and errors from this imperfect correction could become large at low potentials. The authors also studied the relative humidity (RH) effects on ORR kinetics in PEMFCs [32]. They found that when RH is above 50–60%, the kinetics are independent of the RH, but they observed significant ORR kinetic losses at lower RH. The reduction of ORR kinetics at low RH was interpreted as most likely

due to a lowering of the proton activity (therefore only indirectly related to the lowering of the water activity) via hydration of acidic groups and the sequestering of protons at low RH.

Another factor that plays an important role in determining the minimum loading of Pt catalyst required for PEMFCs is the Pt size effect on ORR, not only through altering the fraction of surface Pt atoms over the number of total Pt atoms, but also through changing the ORR kinetics per surface Pt atom (Pt specific activity). Earlier results by Blurton et al. [33] showed that in 20% H<sub>2</sub>SO<sub>4</sub> at 70°C, the highly dispersed Pt (with size of about 1.4 nm) supported on conductive carbon prepared through ion exchange on resin followed by pyrolysis has a Pt-specific activity 20 times lower than that of crystalline Pt black (with size of about 10 nm), though it is not clear to what extent the lower activity could be caused by contamination of Pt and/or a “burying effect” during the catalyst preparation. Blurton et al. [33] correlated the decreased ORR specific activity with the decreased coordination number of surface Pt atoms on smaller Pt particles, causing more severe oxidation of the Pt surface. Peuckert et al. [34] later investigated a series of Pt-on-carbon catalysts with Pt weight percents from 5% to 30% prepared by an impregnation method, with corresponding fractions of metal atoms on the surface from 1.0 (Pt size < 1 nm) to 0.09 (Pt size ~ 12 nm). The Pt-specific activity toward ORR in 0.5 M H<sub>2</sub>SO<sub>4</sub> at 298 K on these catalysts was found to be constant for Pt particle sizes above 4 nm but to decrease by a factor of 20 as the particle size decreased from 3 to 1 nm. Taking into account the larger percentage of buried and these inactive Pt atoms in larger particles, this result suggested that the optimal Pt size for the maximum Pt mass activity is about 3 nm. Kinoshita [35] also reviewed and analyzed the particle size effect for ORR on Pt/C catalysts. Based on the literature data of ORR on Pt/C collected in H<sub>3</sub>PO<sub>4</sub> [36–38] and H<sub>2</sub>SO<sub>4</sub> [34] solutions, Kinoshita proposed the decrease of Pt-specific activity with decrease of Pt particle size is a consequence of the changing distribution of surface atoms at the (100) and (111) crystal faces. Recent literature data [18] in H<sub>3</sub>PO<sub>4</sub> reported that when Pt particle size increases from 2.5 to 12 nm, there is about threefold of increase in Pt-specific activity and confirmed the optimal Pt particle size for maximum mass activity to

be around 3 nm. Gasteiger et al. [21] investigated Pt/C and Pt black catalysts for ORR in HClO<sub>4</sub> solution at 60°C, with the Pt particle size ranging from 2 nm to over 10 nm, and found that the magnitude of activity improvement is comparable to that in literature data [39, 40], although the absolute values are about ten times higher than those reported in Ref. [40], due in part to the use of a less-strongly adsorbing electrolyte.

It is well established that ORR on Pt single crystals is structure-sensitive, depending on the electrolyte. In H<sub>2</sub>SO<sub>4</sub>, the order of activity of Pt(hkl) increases in the sequence (111) << (100) < (110) [19, 41]. The variation in H<sub>2</sub>SO<sub>4</sub> originates from highly structure-specific adsorption of sulfate/bisulfate anions in this electrolyte, which has a strongly inhibiting effect on the (111) surface. Given that the dominant Pt crystal facets of high-surface-area Pt/C catalysts are {111} and {100}, the Pt size effect on ORR activity in H<sub>2</sub>SO<sub>4</sub> can be explained by the structure-sensitive adsorption. In perchloric acid, the variation in activity at is relatively small between the three low index faces, with activity increasing in the order (100) < (111) (110) [42], owing to the structure sensitive inhibiting effect of OH<sub>ads</sub>, i.e., a stronger inhibiting effect on (100) and smaller effects on (110) and (111) [41]. Norskov and other researchers [43–48], based on their DFT calculations, recently proposed using the concept of averaged d-band center energy to explain the reactivity of metal surface atoms, which was supported by numerous experimental data. According to them, when Pt particle size decreases, the average coordination number of surface Pt atoms decreases, causing the d-band center to move closer to the Fermi level and the reactivity of those atoms to increase. As a result of that, the Pt atoms bind oxygen–hydroxide stronger, and therefore, have a lower ORR activity. A stronger adsorption of OH species on Pt surface when the particle size is reduced to below 5 nm was reported in Ref. [49].

While the Pt particle-size effect on ORR suggests that a further increase of Pt dispersion by decreasing the particle size to much smaller than 3 nm will not improve Pt mass activity, it is worth noting that the Pt-size effect is not universally believed. Recently, Yano et al. [50] studied ORR on a series of carbon-supported Pt-nanoparticle electrocatalysts (Pt/C) with average diameters in the range of roughly 1–5 nm, combined with measurements on <sup>195</sup>Pt electrochemical

nuclear magnetic resonance (EC-NMR) spectroscopy. They observed that ORR rate constants and  $\text{H}_2\text{O}_2$  yields evaluated from hydrodynamic voltammograms measurements did not show any particle size dependency. The apparent activation energy of  $37 \text{ kJ mol}^{-1}$ , obtained for the ORR rate constant, was found to be identical to that obtained for bulk platinum electrodes. This was consistent with the negligible difference in the surface electronic properties of these Pt/C catalysts, revealed from the practically no change of surface peak position of  $^{195}\text{Pt}$  NMR spectra and the spin-lattice relaxation time of surface platinum atoms with the particle size variation.

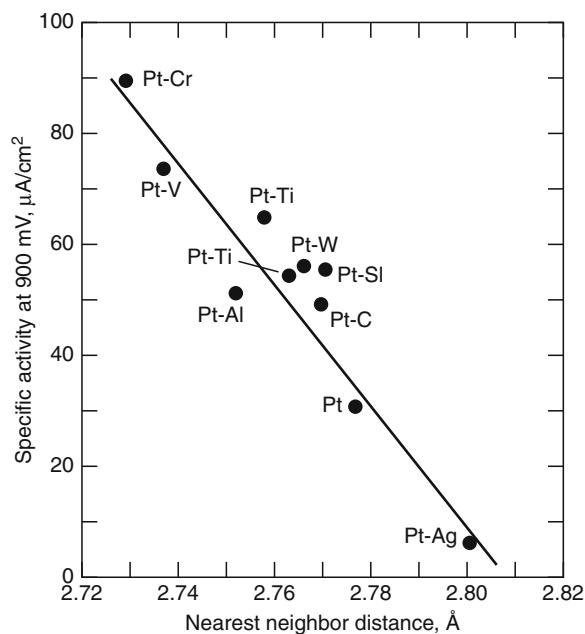
Nevertheless, facilitating the ORR kinetics by augmenting the Pt specific activity is important for the future of fuel cells. A recent perspective in *Science* by Gasteiger and Markovic [51] showed some promising advances in this aspect.

### Pt-Alloy Electrocatalysts

Pt-alloy catalysts, predominantly Pt binary and ternary alloys with 3D transition metals, have been the main focus of catalyst research for PEM fuel cells in the past decades, as they confer enhanced ORR activities over those available from pure Pt catalysts. Great progress has been made in past decades in developing more active and durable Pt alloy catalysts and in understanding the mechanism of their activity enhancements. Two- to threefold specific activity enhancements versus pure Pt were typically reported in literature [21, 40, 52–54], while exceptions existed from earlier results, claiming either no enhancement [55–57], or over an order of magnitude enhancement [58]. As far as which alloy and what alloy compositions confer the highest ORR activity, there seems to be lack of general agreement. This is probably because the measured activity depends highly on the catalyst surface and near-surface atomic composition and structure, on impurities on the surface, and on particle size and shape, all of which could be affected by the preparation method, heat treatment protocol, and testing conditions. For example, to achieve the optimal alloy structure for maximum activity, different Pt alloy particles may require different annealing-temperature protocols to accommodate the distinctions between metal melting points and particle sizes [59, 60]. In a number of earlier

papers [61–63] it was showed that Pt-Cr is the most active ORR cathode catalyst in phosphoric acid fuel cells, while some recent results reported that in the PEMFC-oriented settings, the most active Pt alloy ORR catalyst could be Pt-Co [64], Pt-Fe [58, 65], Pt-Cr [52], Pt-Ni [66], or Pt-Cu [67] at specific atomic ratios of Pt to the alloying elements. Several representative mechanisms have been proposed in the literature to explain the enhanced activities observed on Pt alloy catalysts: (1) a surface roughening effect due to leaching of the alloy base metal [68, 69]; (2) decreased lattice spacing of Pt atoms due to alloying [52, 61, 70]; (3) electronic effects of the neighboring atoms on Pt, such as increased Pt d-band vacancy [52, 58, 65, 71] or depressed d-band center energy upon alloying [42, 64, 67, 72, 73]; and/or (4) decreased Pt oxide/hydroxide formation at high potential [52, 74–76]. The increased Pt surface roughness alone may help increase Pt mass activity but will not increase the Pt-specific activity. Other mechanisms are correlated with each other, for example, the decreased lattice spacing may affect the electronic structure of Pt atoms, which in turn may inhibit the Pt oxide/hydroxide formation. A more detailed discussion follows.

Jalan and Taloy [61] believed that the nearest-neighbor distance between Pt atoms plays an important role in the ORR, based on the reaction model proposed by Yeager et al. [16], i.e., the rate-determining step being the rupture of O–O bond via various dual site mechanisms. They proposed that the distance between nearest-neighbor atoms on the surface of pure Pt is not ideal for dual site adsorption of  $\text{O}_2$  or “ $\text{HO}_2$ ” and that the introduction of foreign atoms that reduce the Pt nearest-neighbor spacing would result in higher ORR activity. By testing a number of carbon-supported Pt-M alloy catalysts fabricated into gas diffusion electrodes, with various nearest-neighbor Pt atom distances determined from X-ray diffraction, a linear relationship was obtained between the activity and the distance, with Pt-Cr exhibiting highest ORR activity and smallest nearest-neighbor distance, as shown in Fig. 3. While the geometric distance between the neighboring Pt atoms is shorter in alloys, the surface electronic structure of Pt alloys is different from that of pure Pt as well, so it is difficult to separate the two factors. Yet other studies [55, 68, 69] claimed no activity enhancement was observed on Pt-Cr alloy over



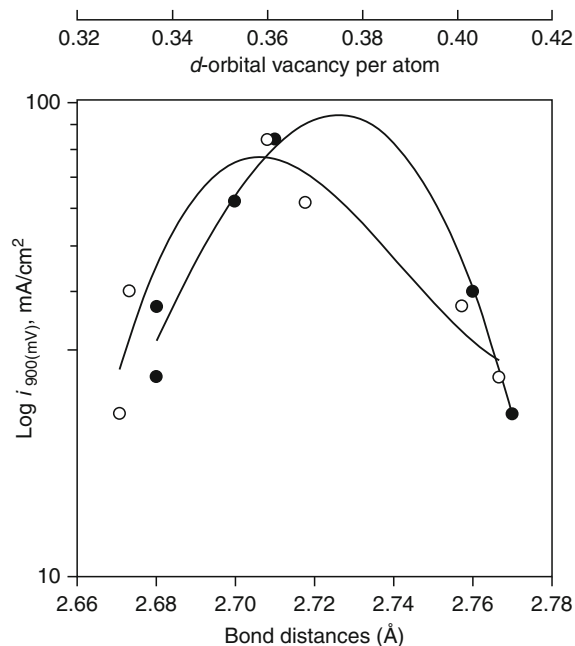
### PEM Fuel Cells and Platinum-Based Electrocatalysts.

**Figure 3**

Specific activity for the ORR versus electrocatalyst nearest-neighbor distance (Reproduced from [61]. With permission)

pure Pt, except for the increased surface roughness [68, 69], but this has not been supported by more recent literature [52, 70].

Mukerjee and coworkers did a series of studies [52, 74, 77, 78] on Pt binary alloys for ORR applying in situ X-ray absorption spectroscopy (XAS) to electrochemical systems. The spectra consist of two parts, the near-edge part XANES (X-ray absorption near-edge structure), which gives chemical information and EXAFS (extended X-ray absorption fine structure), which gives the structural information around the element of interest. In a related study, Mukerjee et al. [52] investigated five binary Pt alloys (PtCr/C, PtMn/C, PtFe/C, PtCo/C, and PtNi/C) supported on high-surface-area carbon for ORR in a proton exchange membrane fuel cell. The electrode kinetic studies on the Pt alloys showed a two- to threefold increase in ORR activity relative to a reference Pt/C electrocatalyst, with the PtCr/C alloy exhibiting the best performance. Contractions in the Pt–Pt bond distances were observed by both EXAFS and XRD. In addition, they found that in the double-layer potential region (0.54 V versus RHE), the alloys possess higher Pt d-band vacancies than Pt/C, while in



### PEM Fuel Cells and Platinum-Based Electrocatalysts.

**Figure 4**

Correlation of the ORR performance of Pt and Pt-alloy electrocatalysts in PEMFC with Pt–Pt bond distance (*solid circles*) and the d-band vacancy of Pt (*empty circles*) obtained from in situ XAS (Reproduced from [52]. With permission)

the high potential region (0.84 V versus RHE), Pt/C shows higher d-band vacancy relative to alloys. This was interpreted by the adsorption of OH species at high potential on Pt/C but to a lesser extent on Pt alloys. Correlation of the electronic (Pt d-band vacancies) and geometric (Pt–Pt bond distance) with the electrochemical performance characteristics exhibits a volcano-type behavior with the PtCr/C alloy being at the top of the curve, shown in Fig. 4. They rationalized the enhanced activity on alloys on the basis of electronic and geometric effects, and of inhibition of OH adsorption. Similar observations were also reported from the ORR measurements in phosphoric acid [66], and in alkaline solution [79–81].

Surface segregation of Pt has been experimentally observed in a wide range of Pt–M binary alloys, such as Pt–Fe [82], Pt–Ni [83], Pt–Co [84], Pt–Ru [85], and has also been reported in theoretical calculations [86, 87]. Furthermore, it has been reported that the topmost layer is composed of pure Pt while the second layer is



enriched in the transition metal M for Pt-rich Fe, Co, and Ni alloys [72, 76, 88], produced by displacement of Pt and M atoms in the first two layers to minimize the surface free energy during annealing. Stamenkovic et al. [76] studied polycrystalline Pt<sub>3</sub>Ni and Pt<sub>3</sub>Co alloys for electrocatalysis of ORR in acid electrolytes using the rotating ring disk electrode (RRDE) method. Polycrystalline bulk alloys of Pt<sub>3</sub>Ni and Pt<sub>3</sub>Co were prepared in ultra-high vacuum (UHV) having two different surface compositions: one with 75% Pt (by sputtering) and the other with 100% Pt (by annealing). The latter was called a “Pt-skin” structure and is produced by an exchange of Pt and Co in the first two layers. Activities of Pt-alloys for the ORR were compared to those of polycrystalline Pt in 0.5 M H<sub>2</sub>SO<sub>4</sub> and 0.1 M HClO<sub>4</sub> electrolytes. It was found that in H<sub>2</sub>SO<sub>4</sub>, the activity increased in the order Pt<sub>3</sub>Ni > Pt<sub>3</sub>Co > Pt; in HClO<sub>4</sub>, however, the order of activities was “Pt-skin/Pt<sub>3</sub>Co” > Pt<sub>3</sub>Co > Pt<sub>3</sub>Ni > Pt. The catalytic enhancement was greater in 0.1 M HClO<sub>4</sub> than in 0.5 M H<sub>2</sub>SO<sub>4</sub>, with the maximum enhancement observed for the “Pt-skin” on Pt<sub>3</sub>Co in 0.1 M HClO<sub>4</sub> being three to four times that for pure Pt. The activity enhancement was attributed to the inhibited Pt-OH<sub>ad</sub> formation on an alloy, even one covered with pure Pt, relative to the surface of a pure-Pt electrode, and the ORR reaction mechanism (pathway) was found to be the same on alloys as on a pure-Pt electrode. In a more recent study, Stamenkovic et al. [64] investigated polycrystalline Pt<sub>3</sub>M (M = Ni, Co, Fe, Ti, V) surfaces for ORR in 0.1 M HClO<sub>4</sub> for both “Pt-skin” and sputtered alloy surfaces. The activity was correlated to the d-band center energy obtained in UHV via ultraviolet photoemission spectroscopy (UPS). A “volcano behavior” was revealed with the Pt<sub>3</sub>Co has the highest activity for both “Pt-skin” and sputtered surfaces. The “Pt-skin” surface was found to be more active than sputtered surface again for each Pt<sub>3</sub>M. The electrochemical and post-electrochemical UHV (ultra-high vacuum) surface characterizations revealed that Pt-skin surfaces are stable during and after immersion to an electrolyte. In contrast, all sputtered surfaces formed Pt-skeleton outermost layers due to dissolution of transition metal atoms [89].

By using DFT calculations, Xu et al. [47] studied the adsorption of O and O<sub>2</sub> and the dissociation of O<sub>2</sub> on the (111) faces of ordered Pt<sub>3</sub>Co and Pt<sub>3</sub>Fe alloys and on monolayer Pt skins covering these two alloys.

Results were compared with those calculated for two Pt(111) surfaces, one at the equilibrium lattice constant and the other laterally compressed by 2% to match the strain in the Pt alloys. The absolute magnitudes of the binding energies of O and O<sub>2</sub> follow the same order in the two alloy systems: Pt skin < compressed Pt(111) < Pt(111) < Pt<sub>3</sub>Co(111) or Pt<sub>3</sub>Fe(111). The reduced bonding strength of the compressed Pt(111) and Pt skins for oxygen was rationalized as being due to the shifting of the d-band center increasingly away from the Fermi level. They proposed that an alleviation of poisoning by O and enhanced rates for reactions involving O could be some of the reasons why Pt skins are more active for the ORR.

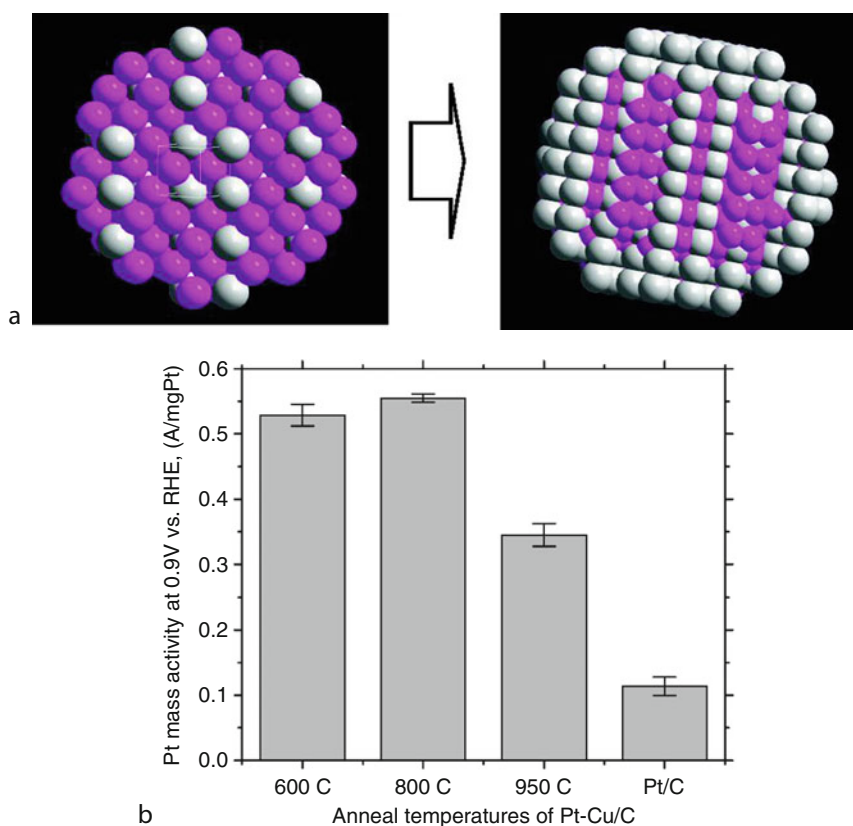
Chen et al. [88, 90] recently studied carbon supported Pt<sub>3</sub>Co nanoparticles for ORR and correlated their activity with the chemical composition and structural information of individual particles. Conventional and aberration-corrected high-angle annular dark-field (HAADF), scanning transmission electron microscopy (STEM), and high resolution transmission electron microscopy (HRTEM) were used to obtain the particle compositional and structural information. For the acid treated Pt<sub>3</sub>Co nanoparticles, they observed the formation of percolated Pt-rich and Pt poor regions within individual nanoparticles, analogous to the skeleton structure observed for sputtered polycrystalline Pt alloy surfaces after acid leaching [89]. The acid treated alloy nanoparticles yielded about two times the specific activity of pure-Pt nanoparticles. After annealing of the acid-treated particles, sandwich-segregation surfaces of ordered Pt<sub>3</sub>Co nanoparticles were directly observed, with the topmost layer being pure Pt atoms. The specific activity of annealed nanoparticles was about four times that of pure Pt nanoparticles. The enhanced Pt-specific activity toward ORR was attributed to the reduced binding energy of oxygenated species, owing to combined two effects, i.e., the increased compressive strain in Pt atoms, and the ligand effect from underlying Co atoms.

Strasser et al. [59, 67, 91–94] recently applied a freeze-drying technique in the synthesis of Pt alloy nanoparticle catalysts with enhanced ORR activity. The Pt-Cu alloy catalyst after electrochemical dealloying was reported to have both mass and specific activities about four to six times those of a standard commercial Pt/C catalyst, in both RDE and MEA tests. The synthesis

involved an impregnation/freeze-drying route followed by annealing. Preparation started with impregnation and sonication of a commercial 30 wt% Pt/C catalyst with an aqueous solution of a copper nitrate, with Pt:Cu atomic ratio of 1:3, followed by freezing in liquid N<sub>2</sub>. The frozen sample was subsequently freeze-dried under a moderate vacuum (0.055 mbar). Reduction and alloying of Pt and Cu on the carbon support was thermally driven under a reductive H<sub>2</sub> atmosphere in a tube furnace. Electrochemical etching (voltammetric dealloying) was employed to remove the surface Cu atoms from Cu-rich Pt-Cu alloy precursors. Bulk and surface structural and compositional characterization suggested that the dealloyed active catalyst phase consists of a core-shell structure in which a multilayer Pt rich shell is surrounding a Pt-poor alloy particle core. This work constitutes significant progress on initial

activity, since a fourfold of increase of Pt mass activity is the performance target for commercially viable fuel cell cathode catalyst [21]. Figure 5a shows a schematic of the dealloying process, and Fig. 5b exhibits the Pt mass activities of Pt-Cu/C synthesized at different temperatures, compared with the Pt/C catalyst.

As to the mechanistic origin of the activity enhancement in dealloyed Pt-Cu catalyst, the authors believe geometric effects play a key role, because the low residual Cu near-surface concentrations make significant electronic interactions between Pt and Cu surface atoms unlikely. Therefore, they suspect that the dealloying creates favorable structural arrangements of Pt atoms at the particle surface, such as more active crystallographic facets or more favorable Pt-Pt interatomic distances for the electroreduction of oxygen, as predicted by DFT calculations [47]. A fourfold



#### PEM Fuel Cells and Platinum-Based Electrocatalysts. Figure 5

(a) The schematic model of a Pt-Cu alloy particle before and after electrochemical dealloying of the near-surface Cu atoms (pink balls = Cu atoms; gray balls = Pt atoms); (b) Pt mass activities of Pt-Cu/C catalysts at various annealing temperatures compared to that of Pt/C catalyst (Reproduced from [67]. With permission)

enhancement in Pt mass activity on monodispersed Pt<sub>3</sub>Co nanoparticles with particle size of 4.5 nm was also reported recently [95].

Watanabe and coworkers carried out a series of studies on the mechanism for the enhancement of ORR activity on Pt-Fe, Pt-Ni, and Pt-Co alloys [53, 58, 65, 71, 96, 97]. By using X-ray photoelectron spectroscopy combined with an electrochemical cell (EC-XPS) [96, 97], they identified quantitatively oxygen-containing species adsorbed on electrodes of a pure Pt and a Pt skin layer (generated by acid treatment, not annealing, and therefore equivalent to the “skeleton” layers described by Stamenkovic et al. [89]) formed on Pt-Fe and Pt-Co alloys' surface from N<sub>2</sub>- and O<sub>2</sub>-saturated 0.1 M HF solution. Four types of species were distinguished with binding energies at 529.6, 530.5, 531.1, and 532.6 eV; the first two were assigned to O<sub>ad</sub> and OH<sub>ad</sub>, while the latter two were assigned to the bilayer water molecules, H<sub>2</sub>O<sub>ad,1</sub> and H<sub>2</sub>O<sub>ad,2</sub>. The XPS results showed that the Pt skin layer exhibited a higher affinity to O<sub>ad</sub> but less to H<sub>2</sub>O compared to pure Pt, particularly in the O<sub>2</sub>-saturated solution. The enhanced ORR activity at the Pt skin/Pt-alloy electrode was ascribed to higher coverage of O<sub>ad</sub> than that at pure Pt. They also found that such an enhancement is induced without changing the activation energy but the corresponding frequency factor value in the pre-exponential term from that of pure Pt. From the measurements in a flow channel with 0.1 M HClO<sub>4</sub>, in the temperature range of 20–50°C, a two to fourfold of ORR-specific activity enhancement was reported for those Pt alloys [53].

Pt metal dissolution in fuel cells has been reported to play a key role in Pt surface area loss and cell performance loss [98]. While Pt-alloy catalysts with fourfold enhancement in both Pt mass activity and specific activity relative to standard Pt/C catalyst seem achievable, as has been shown above, the long-term durability of the alloy catalysts is still a concern, due to dissolution of the base metal from the alloys [21, 99]. For Pt-Cu alloy catalysts, an additional possible risk is that the dissolved Cu may migrate from cathode through membrane to anode and deposit on the anode Pt surface, causing a poisoning effect on anode hydrogen oxidation kinetics, as the Cu redox potential is higher than the anode hydrogen redox potential. Fortunately, the durability of the cathode catalyst can be compensated

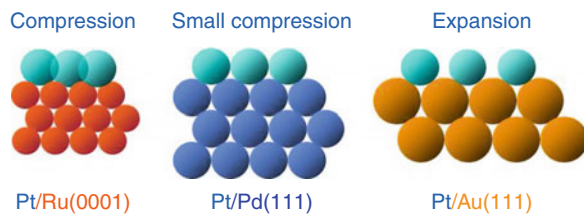
by careful system control, mainly through lowering the cathode upper potential limit and narrowing its operating potential window [100, 101].

### Pt Monolayer Electrocatalysts

Pioneered by Adzic et al. [102, 103], the idea of a Pt monolayer electrocatalyst has been one of the key concepts in reducing the Pt loading of PEM fuel cells in recent years. Pt submonolayers deposited on Ru nanoparticles had been earlier demonstrated to give superior performance with ultra-low Pt loading compared to commercial Pt/C or Pt-Ru alloy catalysts for the anode CO-tolerant hydrogen oxidation reaction [103–107]. More recently, Adzic and coworkers applied this concept in making novel Pt monolayer catalysts for the cathode ORR, which will be the focus of the review in this section. In general, the new method of synthesizing Pt monolayer catalysts involves underpotential deposition (UPD), a technique well known to produce an ordered atomic monolayer to multilayer metal deposition onto a foreign metal substrate [108–110]. Specifically, the method consists of two steps [111, 112]: first, a monolayer of a sacrificial less-noble metal is deposited on a more noble metal substrate by UPD, such as Cu UPD on Au or Pd; second, the sacrificial metal is spontaneously and irreversibly oxidized and dissolved by a noble metal cation, such as a Pt cation, which is simultaneously reduced and deposited onto the foreign metal substrate. The whole procedure can be repeated in order to deposit multilayers of Pt (or another noble metal) on the foreign metal.

The advantages of Pt monolayer catalysts include (1) full utilization of the Pt atoms that are all on the surface, and (2) that the Pt activity and stability can be tailored by the selection of the substrate metals. For example [102], when a Pt monolayer is deposited onto different substrate metals, as shown in Fig. 6, due to the lattice mismatch between the metals, it can experience compressive or tensile stress, which is known to affect the Pt activity by adjusting its d-band center energy [43, 47] and consequently its ORR activity.

Pt monolayer deposits on Pd(111) single crystals (Pt/Pd(111)) and on Pd/C nanoparticles (Pt/Pd/C) have been studied for ORR [112], and improved activities compared to Pt(111) and commercial Pt/C, respectively, were reported. The ORR reaction mechanism of



### PEM Fuel Cells and Platinum-Based Electrocatalysts.

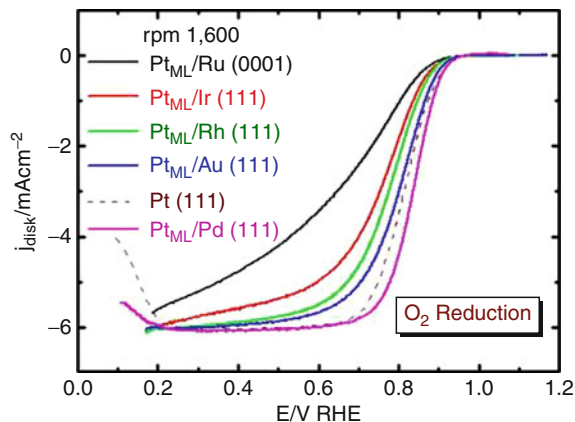
**Figure 6**

Models of pseudomorphic monolayers of Pt on three different substrates inducing compressive strain (Ru(0001) and Pd(111)) and tensile strain (Au(111)). (Reproduced from [102]. With permission)

the monolayer catalysts was found to be the same as that on pure Pt surface. Pt/Pd(111) was found to have a 20 mV improvement in half-wave potential versus Pt(111), and the Pt/Pd/C had a Pt-mass activity five- to eight-times higher than that of Pt/C catalyst. If the total noble metal amount (Pt + Pd) is counted, the mass activity is about 80% higher than that of Pt/C catalyst [112]. The enhanced ORR activity is attributed to the inhibited OH formation at high potential, as evidenced from XAS measurements. In a real fuel cell test, 0.47 g Pt/kW was demonstrated at 0.602 V [113].

In order to understand the mechanism for the enhanced activity of a Pt monolayer deposited on Pd metal, the ORR was investigated in O<sub>2</sub>-saturated 0.1 M HClO<sub>4</sub> solution on platinum monolayers supported on Au(111), Ir(111), Pd(111), Rh(111), and Ru(0001) single-crystal RDE surfaces [114]. A comparison of the polarization curves at 1,600 rpm is shown as in Fig. 7. The trend of the ORR activities increase in the sequence Pt/Ru(0001) < Pt/Ir(111) < Pt/Rh(111) < Pt/Au(111) < Pt(111) < Pt/Pd(111).

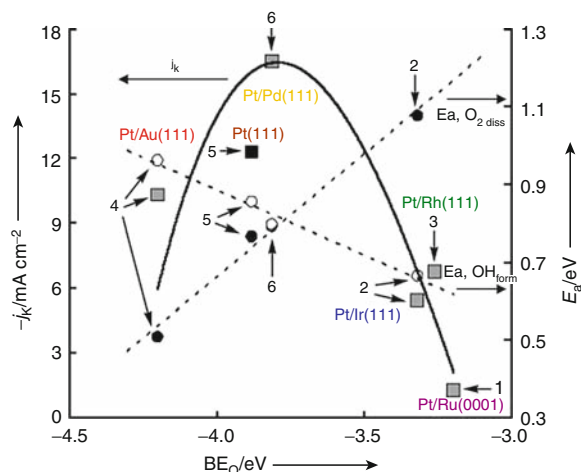
The authors further correlated the kinetic activities of the “monolayers” with the Pt d-band center energies (not shown) and Pt-O binding energies, and found a “volcano” relationship, with the Pt/Pd(111) having the optimal d-band center, as well as Pt<sub>ML</sub>-O-binding energy, for the maximum ORR activity. The “volcano” behavior was rationalized as being controlled by the two key steps in ORR, the O–O bond dissociation which is followed by the other step, the O–H bond formation. As shown in Fig. 8, the activation energies of the two steps both correlated linearly with the Pt<sub>ML</sub>-O-binding energies (and with the



### PEM Fuel Cells and Platinum-Based Electrocatalysts.

**Figure 7**

Polarization curves for O<sub>2</sub> reduction on platinum monolayers in 0.1 M HClO<sub>4</sub> solution on a RDE. Rotation rate = 1,600 rpm (Reproduced from [114]. With permission)



### PEM Fuel Cells and Platinum-Based Electrocatalysts.

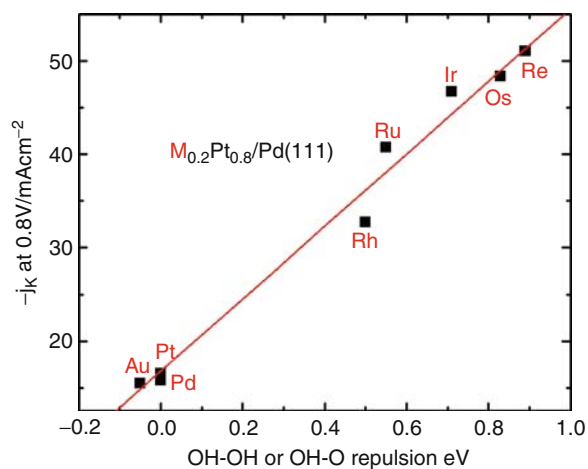
**Figure 8**

Kinetic currents ( $j_k$ ; square symbols) at 0.8 V (versus RHE) calculated from Fig. 7 for ORR and the activation energies for O<sub>2</sub> dissociation (solid circles) and for OH formation (empty circles) on Pt/Ru(0001) (1); Pt/Ir(111) (2), Pt/Rh(111) (3), Pt/Au(111) (4), Pt(111) (5) and Pt/Pd(111) (6), as functions of the calculated binding energy of atomic oxygen (BEO) (Reproduced from [114]. With permission)

d-band center energies, not shown), but in the opposite trend, indicating the Pt<sub>ML</sub>-O binding can be neither too strong, nor too weak, for the best ORR activity.

For further fine-tuning of the monolayer Pt/Pd ORR activity, they further introduced mixed metal + Pt monolayer catalysts [115], which contained 0.2 monolayer of a foreign metal from selection of (Au, Pd, Rh, Ir, Ru, Os, and Re) combined with 0.8 monolayer of Pt co-deposited on Pd(111) or on Pd/C nanoparticles. The foreign metals have either a weaker M–OH bond (for the case of Au–OH), or a stronger M–OH bond (for the rest of the cases) than the Pt–OH bond. DFT calculations [115] showed that, in addition to altering the Pt d-band center energies, the  $\text{OH}_{(-M)}-\text{OH}_{(-Pt)}$  (or  $\text{O}_{(-M)}-\text{OH}_{(-Pt)}$ ) repulsion plays an important role in augmenting the ORR activity, as shown below in Fig. 9. Instead of adjusting the composition of the top-most Pt monolayer, replacing the substrate Pd(111) with Pd<sub>3</sub>Fe(111) to generate Pt/Pd<sub>3</sub>Fe(111) was recently reported to also have an enhanced ORR activity [116].

Another type of Pt monolayer catalyst showing improved ORR activities are Pt monolayers deposited on (noble metal)/(non-noble metal) core-shell nanoparticles [117]. The synthesis approach started with impregnation of high surface carbon into a mixed solution of noble metal precursor and non-noble metal precursor, followed by stir-drying in air. Core-shell metal substrates were formed by surface



#### PEM Fuel Cells and Platinum-Based Electrocatalysts.

**Figure 9**

Kinetic current at 0.80 V as a function of the calculated interaction energy between two OHs, or OH and O. Positive energies indicate more repulsive interaction compared to Pt/Pd(111) (Reproduced from [115]. With permission)

segregation of the noble metal at elevated temperature in a reductive atmosphere. A Pt monolayer was then deposited on the core/shell substrates by galvanic displacement of a Cu monolayer that was UPD-deposited onto the core-shell substrate particles. Three combinations were investigated: Pt/Au/Ni, Pt/Pd/Co, and Pt/Pt/Co. The enhancement of Pt mass activity of the best case (Pt/Au/Ni) was reported as being over an order of magnitude relative to the commercial Pt/C catalyst. The total noble metal mass activities were reported to be 2.5–4 times higher than that of Pt/C, with the Pt/Pt/Co having the highest number. The enhancement of activities was attributed to the geometric effect induced by the fine-tuning of the Pt lattice spacing with the substrate core-shell particles and to the inhibited PtOH formation because of lowering of the d-band center position relative to the Fermi level.

In a related study, Zhang et al. [118] investigated a partial monolayer of Au deposited on Pt(111) and on Pt/C nanoparticles for ORR. The catalysts were synthesized by Au displacement of a monolayer Cu that was deposited with UPD onto the Pt(111) or Pt/C substrates. Due to the valence-state difference between  $\text{Au}^{3+}$  and  $\text{Cu}^{2+}$ , two thirds of a monolayer of Au was deposited onto each of the Pt surfaces. The Au atoms appeared to form clusters on Pt surfaces. While the ORR activities of the Au/Pt(111) and Au/Pt/C catalysts were slightly lower than those of pure Pt(111) and Pt/C catalysts, respectively; the stability of the Au/Pt/C was superior compared to that of Pt/C catalyst. Potential cycling tests were performed on RDE in 0.1 M  $\text{HClO}_4$  solution, with the potential cycling window between 0.6 and 1.1 V, for 30,000 cycles. The catalytic activity of Au/Pt/C, measured as half-wave potentials on the  $\text{O}_2$  reduction polarization curves obtained before and after potential cycling, showed only a 5 mV degradation in over the cycling period. In contrast, the corresponding change for Pt/C amounted to a loss of 39 mV. In situ X-ray absorption near-edge spectroscopy (XANES) with respect to the potential applied on the catalyst surfaces revealed the oxidation of Pt nanoparticles covered by Au at high potentials was decreased in comparison with the oxidation of Pt nanoparticles lacking such coverage.

More recently, Wang et al. [119] investigated the ORR on well-defined Pt/Pd and Pt/PdCo core-shell nanoparticles for the effects of particle size, facet

orientation, and Pt shell thickness. The Pt shell was generated and the shell thickness was controlled by a novel method called the Cu-UPD-mediated electro-deposition method, in which repeated Cu-UPD/stripping, potential cycling, and Pt irreversible deposition occurred simultaneously in the same electrolyte, with the Pt deposition under diffusion control, until a desired thickness of Pt was achieved. Atomic level analysis using Z-contrast scanning transmission electron microscopy (STEM) coupled with element-sensitive electron energy loss spectroscopy (EELS) showed that well-controlled core-shell particles were obtained. ORR tests on RDE showed that Pt-monolayer catalysts on 4 nm Pd and 4.6 nm Pd<sub>3</sub>Co cores exhibited 1.0 and 1.6 A/mg Pt mass activities at 0.9 V, respectively, about five- and ninefold enhancements over that of 3 nm Pt nanoparticles. Also, two- and threefold enhancements in specific activity were observed respectively, which were mainly attributed to the nanosize- and lattice-mismatch-induced contraction in (111) facets based on the DFT calculations using a nanoparticle model. Scale-up methods were developed for synthesis of the core-shell particles [120].

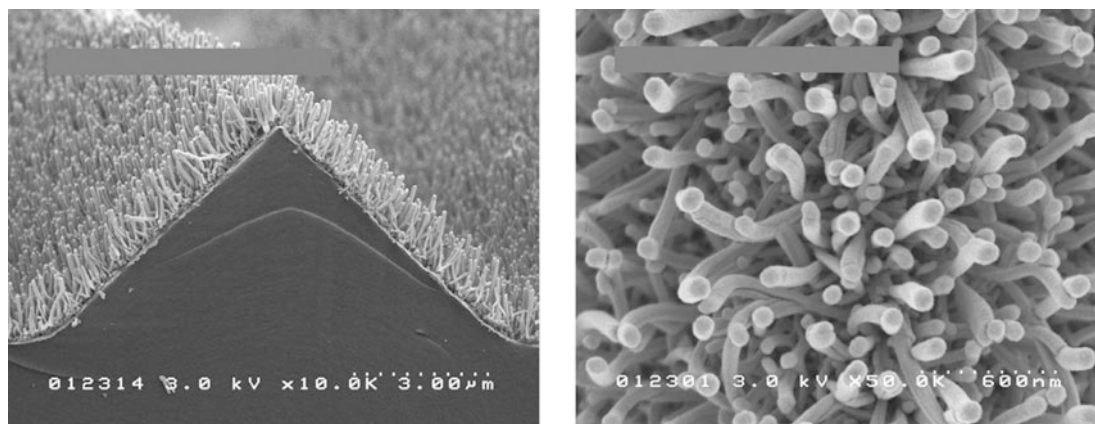
In summary, Pt monolayer catalysts show a promising pathway toward solving one of the major problems facing PEM fuel cells by enhancing the Pt-specific activity and the utilization of Pt atoms, and therefore reducing the cost of the cathode catalyst, although more fuel cell tests of durability are needed before the monolayer catalysts can be put in fuel cell vehicles. There is still a need for a reduction in the total noble metals in these catalysts.

### Pt and Pt-Alloy Nanowire and Nanotube Electrocatalysts

As has been discussed in the introduction, low activity and limited durability are the two major issues related to the high-surface-area-carbon-dispersed Pt nanoparticle catalysts (Pt/C) in PEM fuel cells. Shao-Horn et al. [121] have given a comprehensive review on the instability considerations of Pt/C catalysts. Conventional PEM fuel cell catalysts typically consist of Pt nanoparticles in size of 2–3 nm that are supported on high-surface carbon with 20–50 nm primary carbon particles for electrical conductivity and high levels of catalyst activity. As it has been shown above, this kind

of catalyst has already approached the maximum Pt mass activity as a pure Pt/C catalyst, as the Pt particle size is close to the optimal value. However, the 2–3 nm Pt nanoparticles are intrinsically not stable enough under PEM fuel cell operating conditions if no system-mitigation methods are applied. As an example, for a fuel cell short stack operating at steady-state open-circuit voltage (OCV,  $\sim 0.95$  V versus RHE), with H<sub>2</sub>/air flows (stoichiometric reactant flows of  $s = 2/2$ ) at 80°C, fully humidified and 150 kPa<sub>abs</sub> for 2,000 h, the Pt surface area decreased from  $\sim 70$  m<sup>2</sup>/g Pt to  $\sim 15$  m<sup>2</sup>/g Pt, corresponding to an almost 80% of Pt surface area loss [98]. If accompanied by a similar percentage loss of activity, this is not acceptable for commercialization of fuel cell vehicles. In addition, corrosion of carbon support makes the situation even more serious [122, 123]. One way to address this issue is to lower the upper voltage limit of the fuel cell through system mitigation [100, 101], or alternatively, to improve the intrinsic Pt stability through catalyst design, such as using Pt or Pt-alloy nanowire/nanotube catalysts. The local curvature of the nanowire/nanotube Pt or its alloy is expected to be small (in at least one direction), and it consequently has a lower surface free energy and higher stability. On the other hand, the Pt-specific activity of the nanowire/nanotube is expected to be higher than its nanoparticle counterpart, as it should bind OH/O less strongly.

The 3 M nanostructured thin film (NSTF) catalyst is such a non-conventional catalyst [124, 125]. The support particle is a crystalline organic pigment material, perylene red (PR), which is vacuum-deposited and converted to an oriented whisker phase by thermal annealing. The result is a uniquely structured thin film composed of highly oriented, densely packed crystalline organic whiskers [126]. A crystalline Pt coating can be vacuum-deposited on the whiskers. Figure 10 shows SEM images of the NSTF catalyst-coated whiskers prior to incorporation onto the surfaces of a PEM to form a catalyst-coated membrane [127]. The cross section of the whiskers is on the order of 50 nm, and the lengths of the whiskers are controllable by the thickness of the as-deposited PR film, typically in the range of 0.5–2  $\mu$ m. For a practical loading of  $\sim 0.2$  mg Pt/cm<sup>2</sup>, the typical Pt crystallite size of the coated PR whiskers is 10–11 nm, and the specific surface areas of the NSTF-Pt catalysts are  $\sim 10$  m<sup>2</sup>/g Pt [124].

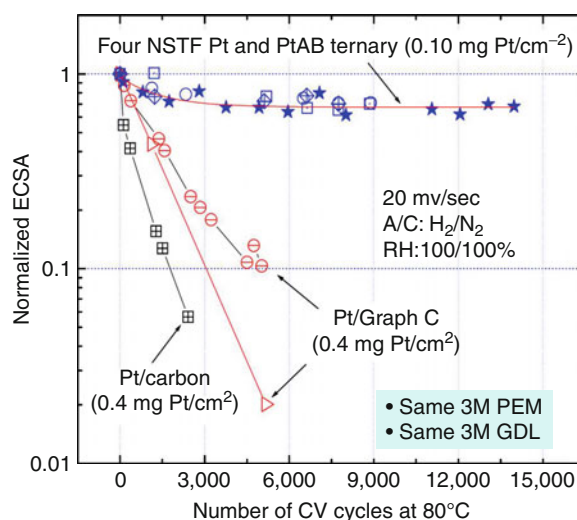


**PEM Fuel Cells and Platinum-Based Electrocatalysts. Figure 10**

Scanning electron micrographs of typical NSTF catalysts as fabricated on a microstructured catalyst transfer substrate, seen (*top*) in cross section with original magnification of  $\times 10,000$ , and (*bottom*) in plain view with original magnification of  $\times 50,000$ . The dotted scale bar is shown in each micrograph (Reproduced from [127]. With permission)

Bonakdarpour et al. [128] investigated Pt/NSTF and Pt-Co-Mn/NSTF for ORR by using the rotating ring disk electrode (RRDE). The nominal chemical composition of the ternary alloy was calculated to be  $\text{Pt}_{0.68}\text{Co}_{0.3}\text{Mn}_{0.02}$ . The catalyst-coated whiskers were carefully brushed off of the original substrate web and applied onto the glassy carbon disk of the RRDE. The measurements were done in  $\text{O}_2$ -saturated 0.1 M  $\text{HClO}_4$  at room temperature. The Pt-specific activity of the Pt/NSTF was found to be close to that of a Pt polycrystalline disk. A twofold gain of Pt specific activity was observed on Pt-Co-Mn/NSTF versus Pt/NSTF. In PEMFC measurements for the same loading of 0.2 mg  $\text{Pt}/\text{cm}^2$ , using GM recommended conditions for measuring specific and mass activity at  $80^\circ\text{C}$ , saturated  $\text{H}_2/\text{O}_2$  at 150  $\text{kPa}_{\text{abs}}$ , at 900 mV, the NSTF Pt-Co-Mn catalyst-coated membrane (CCM) generated specific activities of 2.93  $\text{mA}/\text{cm}^2\text{Pt}$ , and mass activities of 0.18  $\text{A}/\text{mg}^1\text{Pt}$ . The specific activity is  $\sim 12$  times higher, and mass activity is about two times higher than those of TKK 47 wt% Pt/C.

To investigate the NSTF electrode stability under high voltage cycling, Debe et al. [129] tested a series of NSTF Pt and NSTF Pt-ternary catalysts along with Pt/C (Ketjen Black) and Pt/C (graphitic) types by scanning at 20 mV/s, between 0.6 and 1.2 V under saturated  $\text{H}_2/\text{N}_2$  at  $80^\circ\text{C}$ . The MEAs contained the same 3 M-ionomer PEM having a 1,000 equivalent weight (EW), and the same 3 M-coated GDL. The NSTF



**PEM Fuel Cells and Platinum-Based Electrocatalysts. Figure 11**

Normalized surface area versus number of CV cycles from 0.6 to 1.2 V for four NSTF catalyst samples and three Pt/carbon catalysts at  $80^\circ\text{C}$ . All MEAs used the same 3 M ionomer PEM and GDL (gas diffusion layer) (Reproduced from [129]. With permission)

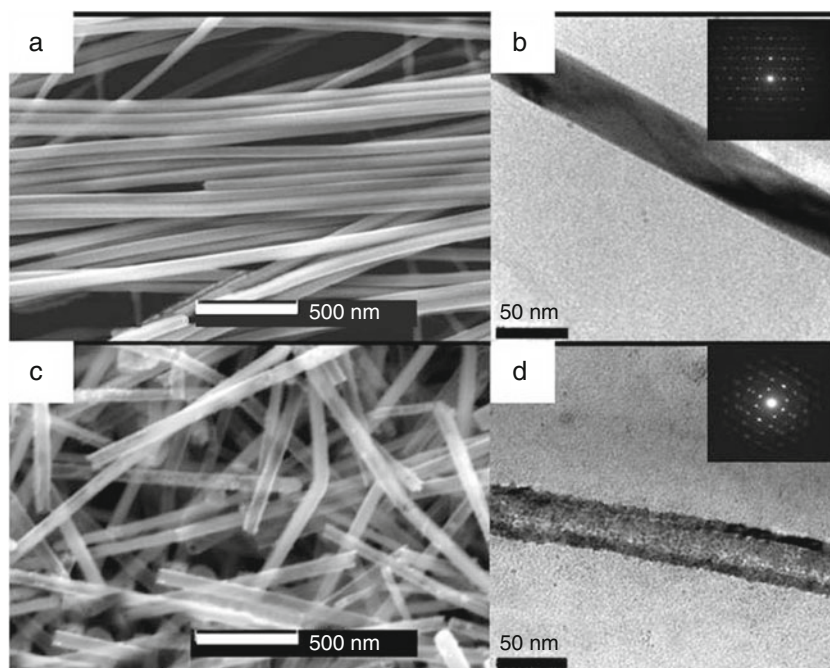
electrodes had loadings of  $0.1\text{ mg}/\text{cm}^2$ , while the carbon- and graphitic-carbon-supported catalysts had loadings of  $0.4\text{ mg}/\text{cm}^2$ . Figure 11 compares the normalized surface area as a function of the number of CV cycles for all the samples. It is interesting to note that

for the NSTF samples the Pt and ternary catalysts behave similarly and lose approximately 30% of the surface area out to 14,000 cycles. The Pt/C and Pt/graphitic carbon, on the other hand, lose substantially more surface area in significantly fewer cycles. The superior ability of the NSTF catalyst to withstand thousands of fast voltage scans over the potential range most critical for Pt dissolution and Pt agglomeration demonstrated a significant differentiating feature over carbon-supported catalysts [129].

Recently, Chen et al. [130] developed supportless Pt nanotubes (PtNTs) and Pt-alloy nanotubes (e.g., Pt-Pd alloy nanotubes (PtPdNTs)) as cathode catalysts for PEMFCs. PtNTs were synthesized by a galvanic replacement of silver nanowires (AgNWs) by following a similar method developed by Xia and coworkers [131, 132]. The AgNWs were synthesized using a polyol method [133] and were subsequently heated at reflux temperature with  $\text{Pt}(\text{CH}_3\text{COO})_2$  in an aqueous solution. For the preparation of PtPdNTs, mixed aqueous  $\text{Pt}(\text{CH}_3\text{COO})_2$  and  $\text{Pd}(\text{NO}_3)_2$  solutions were used. The diameter (Fig. 12a, b) and length of AgNWs

are about 40 nm and 10  $\mu\text{m}$ , respectively. After Pt replacement, the diameter, wall thickness (Fig. 12c, d), and length of the PtNTs are about 40 nm, 6 nm, and 10  $\mu\text{m}$ , respectively.

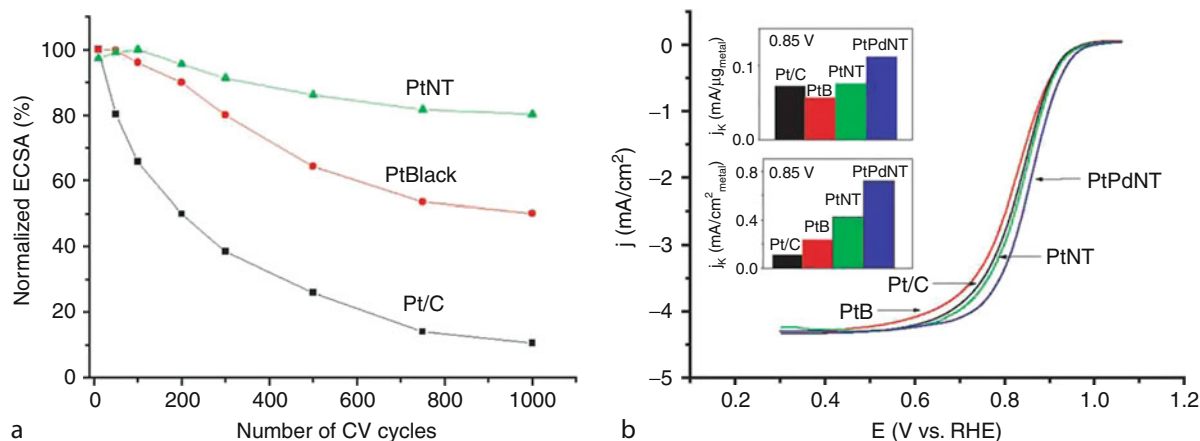
Chen et al. [130] tested the durability of these materials by cycling the electrode between 0 and 1.3 V versus RHE at a scan rate of 50 mV/s in argon-purged 0.5 M  $\text{H}_2\text{SO}_4$  solution at 60°C. As shown in Fig. 13a, the electrochemical surface area (ECSA) of the PtNTs decreased by about 20% after 1,000 cycles, while the Pt-black and Pt/C catalysts lost about 51% and 90% of their platinum ECSA, respectively. Figure 13b shows the comparison of typical ORR polarization curves of the respective catalysts obtained at room temperature in  $\text{O}_2$ -saturated 0.5 M  $\text{H}_2\text{SO}_4$  using a rotating disk electrode (RDE) at 1,600 rpm. At 0.85 V versus RHE, the mass activity of PtNTs was reported slightly higher than that of Pt/C, and 1.4 times higher than that of Pt-black catalysts. The specific activity of the PtNTs was 3.8 times and 1.8 times higher than those of Pt/C and Pt-black catalysts, respectively. For PtPdNTs, the mass activity was measured 1.4 times and 2.1 times higher



PEM Fuel Cells and Platinum-Based Electrocatalysts. Figure 12

(a) SEM image of AgNWs. (b) TEM image and electron diffraction pattern (inset) of AgNWs. (c) SEM image of PtNTs. (d) TEM image and electron diffraction pattern (inset) of PtNTs (Reproduced from [130]. With permission)





### PEM Fuel Cells and Platinum-Based Electrocatalysts. Figure 13

(a) Loss of electrochemical surface area (ECSA) of Pt/C (E-TEK), platinum-black (PtB; E-TEK), and PtNT catalysts with number of potential cycles in Ar-purged 0.5 M H<sub>2</sub>SO<sub>4</sub> solution at 60°C (0–1.3 V versus RHE, sweep rate 50 mV/s). (b) ORR curves (shown as current–voltage relations) of Pt/C, platinum black (PtB), PtNTs, and PdPtNTs in O<sub>2</sub>-saturated 0.5 M H<sub>2</sub>SO<sub>4</sub> solution at room temperature (1,600 rpm, sweep rate 5 mV/s). Inset: Mass activity (*top*) and specific activity (*bottom*) for the four catalysts at 0.85 V (Reproduced from [130]. With permission)

than those of Pt/C and Pt-black catalysts, while the specific activity was 5.8 times and 2.7 times higher than those of Pt/C and Pt-black catalysts, respectively.

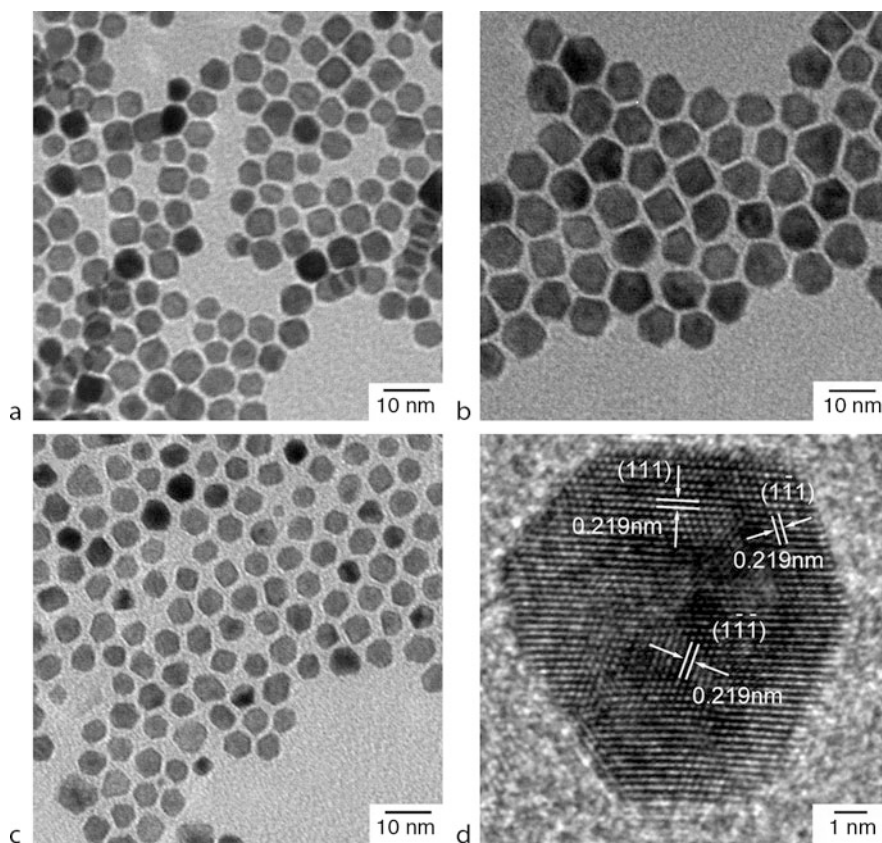
The examples given above illustrate that it is possible to achieve both activity and durability goals of Pt-based catalysts through special morphology and structural design of the catalyst and electrode. There are many other important advancements not discussed here on nanostructured Pt and Pt-alloy catalysts/electrodes showing improvements in activity and/or durability for ORR, such as single-crystal-Pt nanowires grown on continuous carbon-layer-coated Sn-fiber 3D electrodes [134], Pt-Pd bimetallic nanodendrites [135, 136], faceted Pt nanocrystals [137], nanoporous Pt alloy electrodes [138–140].

### Facet- and Shape-Controlled Pt-Alloy Nanocrystal Electrocatalysts

Stamenkovic and coworkers [42] demonstrated in HClO<sub>4</sub> that on RDE Pt<sub>3</sub>Ni(111) single-crystal surfaces of ~6 mm in diameter, the specific activity for ORR is about an order of magnitude higher than on the Pt (111) surface and is about 90 times higher than on Pt/C catalyst, while the other two low-index surfaces, [Pt<sub>3</sub>Ni (100) and Pt<sub>3</sub>Ni(110)] are much less active than

Pt<sub>3</sub>Ni(111). This result is very intriguing in that it suggests that if one can make Pt<sub>3</sub>Ni nanocrystals with all exposed surfaces having {111} orientations, one can hope to gain an enhancement of specific activity by up to two orders of magnitude relative to state-of-the-art Pt/C catalysts. Recently, two interesting papers [141, 142] have shown progress on synthesizing such Pt alloy nanocrystals.

Wu et al. [141] recently reported an approach to the preparation of truncated-octahedral Pt<sub>3</sub>Ni (*t,o*-Pt<sub>3</sub>Ni) catalysts that have dominant exposure of {111} facets. The shape-defined Pt-Ni nanoparticles were made from platinum acetylacetonate [Pt(acac)<sub>2</sub>] and nickel acetylacetonate [Ni(acac)<sub>2</sub>] in diphenyl ether (DPE) using a mixture of borane-tert-butylamine complex (TBAB) and hexadecanediol as the reducing agents. Long-alkane-chain amines were used as the main capping agents, and adamantancarboxylic acid (ACA) or adamantaneacetic acid (AAA) was used to affect the reaction kinetics. The population of truncated octahedral crystals could be adjusted by the types and amounts of the reducing agents and capping agents. Three sets of Pt<sub>3</sub>Ni nanocrystals were generated with various truncated-octahedral crystal populations; see Fig. 14.



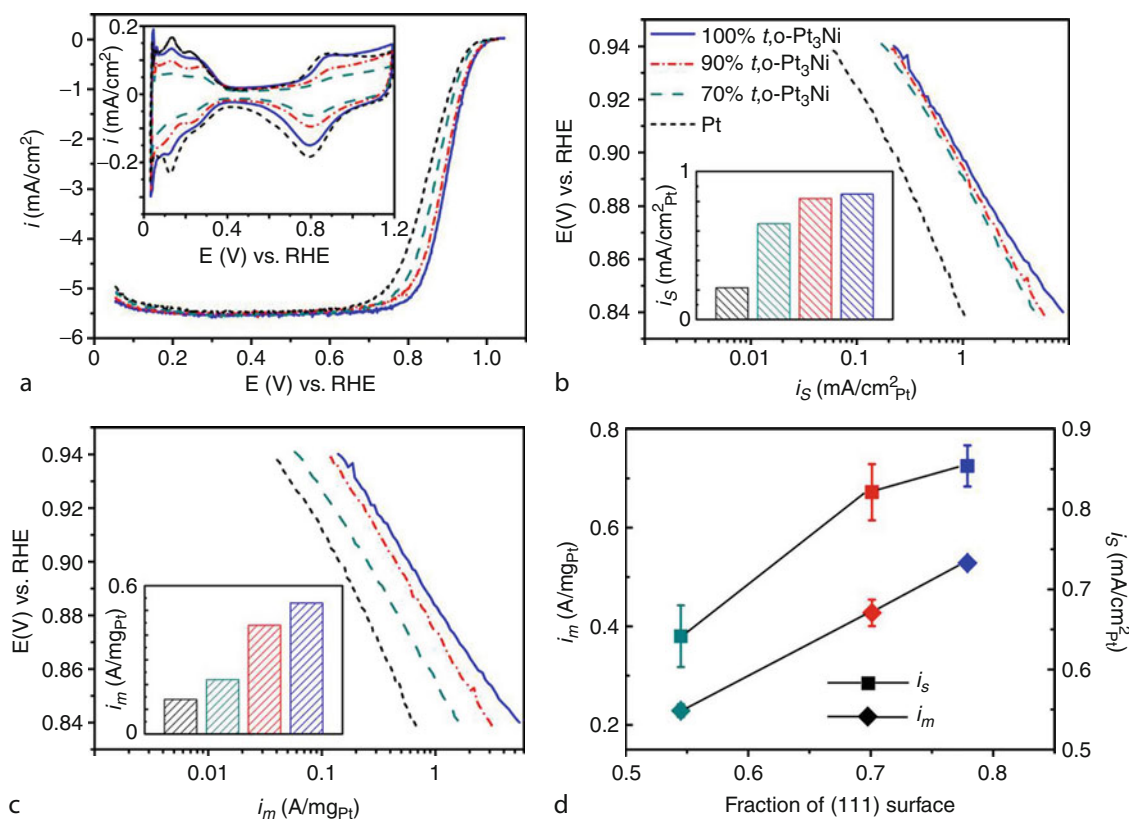
PEM Fuel Cells and Platinum-Based Electrocatalysts. **Figure 14**

TEM images of  $\text{Pt}_3\text{Ni}$  nanocrystals with truncated octahedron population of (a) 70%, (b) 90%, (c) 100%, and (d) HR-TEM image of a truncated octahedron showing the (111) lattice (Reproduced from [141]. With permission)

While Fig. 14a, b contain 30% and 10% of  $\text{Pt}_3\text{Ni}$  cubes (with the remaining particles being truncated-octahedrons), respectively, Fig. 14c contains only truncated-octahedrons. The particle size is on the order of 5 to 7 nm. Only two types of facets are exposed of all the nanocrystals, i.e., the {111} and {100}. The fractions of the {111} surface area over the total surface area could be calculated based on the geometries of the shapes and the population statistics. The ORR kinetics of the nanocrystals were studied on RDEs in  $\text{O}_2$ -saturated 0.1 M  $\text{HClO}_4$ , at room temperature, at 1,600 rpm, with a potential scan rate of 10 mV/s. Figure 15 shows comparison of polarization curves, cyclic-voltammety curves, mass activities, and specific activities of the  $\text{Pt}_3\text{Ni}$  nanocrystals to the standard TTK Pt/Vulcan carbon catalyst. As shown in Fig. 15d, almost-linear correlations were obtained for both mass activities and

specific activities versus the fraction of the (111) surface area over the total surface area. A tabulated kinetic activity comparison is shown in Table 1. The mass activity and specific activity comparisons were made at 0.9 V versus RHE. Roughly  $4\times$  mass-activity and specific-activity enhancements were observed on the 100% truncated octahedral  $\text{Pt}_3\text{Ni}$  nanocrystals over the Pt/C catalyst. While the {111} facets of the nanocrystals showed much higher specific activity than the {100} facets, as indicated in Fig. 15d, in agreement with trend found on bulk  $\text{Pt}_3\text{Ni}$  single crystal disks, the absolute values of the specific activities of the nanocrystals are still far below those observed on bulk single-crystal surfaces [42].

Another interesting report is from Zhang et al. [142] on the synthesis and ORR activity of  $\text{Pt}_3\text{Ni}$  nano-octahedra and nanocubes, with the two shapes



PEM Fuel Cells and Platinum-Based Electrocatalysts. **Figure 15**

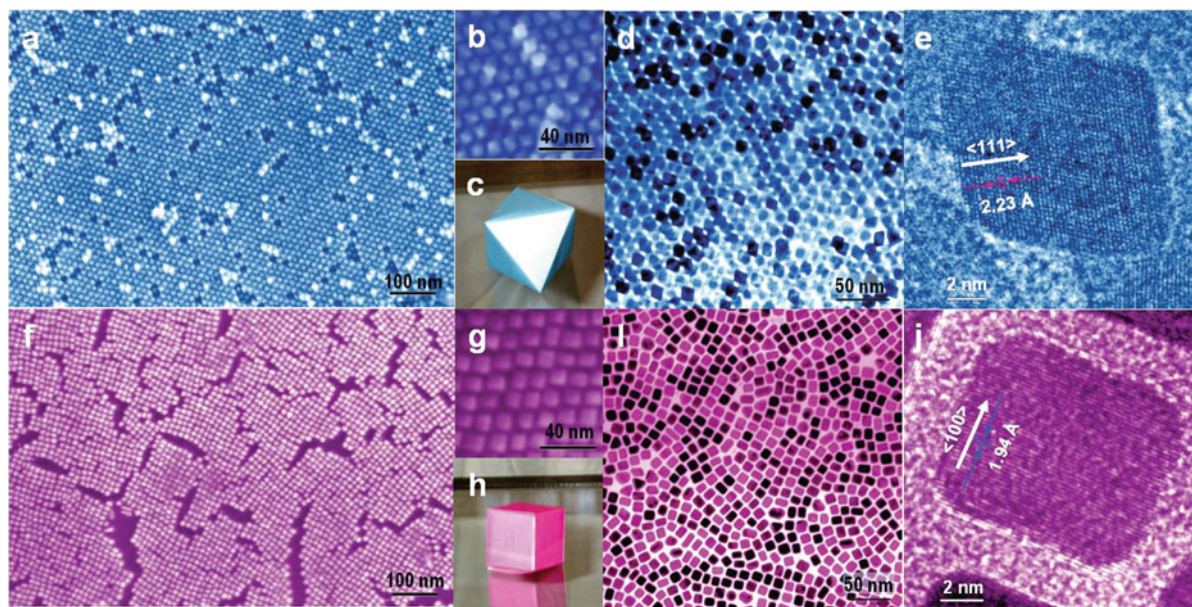
(a) Polarization curves and CV curves (*inset*), (b) area ( $\text{mA}/\text{cm}^2\text{Pt}$ ), (c) mass ( $\text{A}/\text{mg Pt}$ ) specific ORR activity for the *t,o*-Pt<sub>3</sub>Ni and reference Pt catalysts; and (d) the correlations between specific activities and fractions of (111) surfaces of these Pt<sub>3</sub>Ni catalysts. The ORR polarization curves were collected at 1,600 rpm (Reproduced from [141]. With permission)

PEM Fuel Cells and Platinum-Based Electrocatalysts. **Table 1** ECSA, mass- and area- specific ORR activities of Pt<sub>3</sub>Ni and Pt/catalysts (at 0.9 V versus RHE) (Reproduced from [141]. With permission)

Sample name	Pt loading [ $\mu\text{g}_{\text{Pt}}/\text{cm}_{\text{disk}}^2$ ]	ECSA [ $\text{m}^2/\text{g}_{\text{Pt}}$ ]	Mass activity [ $\text{A}/\text{mg}_{\text{Pt}}$ ]	Specific activity [ $\text{mA}/\text{cm}_{\text{Pt}}^2$ ]
100% <i>t,o</i> -Pt <sub>3</sub> Ni	9.3	62.4	0.53	0.85
90% <i>t,o</i> -Pt <sub>3</sub> Ni	9.3	53.7	0.44	0.82
70% <i>t,o</i> -Pt <sub>3</sub> Ni	9.3	33.8	0.22	0.65
Pt/C (TKK)	11	65	0.14	0.215

of nanocrystals having only {111} facets and {100} facets exposed, respectively. The monodispersed Pt<sub>3</sub>Ni nano-octahedra and nanocubes were synthesized via a high-temperature organic solution chemistry approach, which involved using mixed oleylamine and oleic acid at elevated temperature as the reducing

agent and capping agent, and tungsten hexacarbonyl W(CO)<sub>6</sub> as the shape controlling agent. Detailed procedures for synthesis of the nanocrystals can be found in Ref. [142]. Figure 16 shows the SEM and TEM images of those shape and size controlled nano-octahedral (a–e) and nanocube (f–j) crystals. The chemical



PEM Fuel Cells and Platinum-Based Electrocatalysts. Figure 16

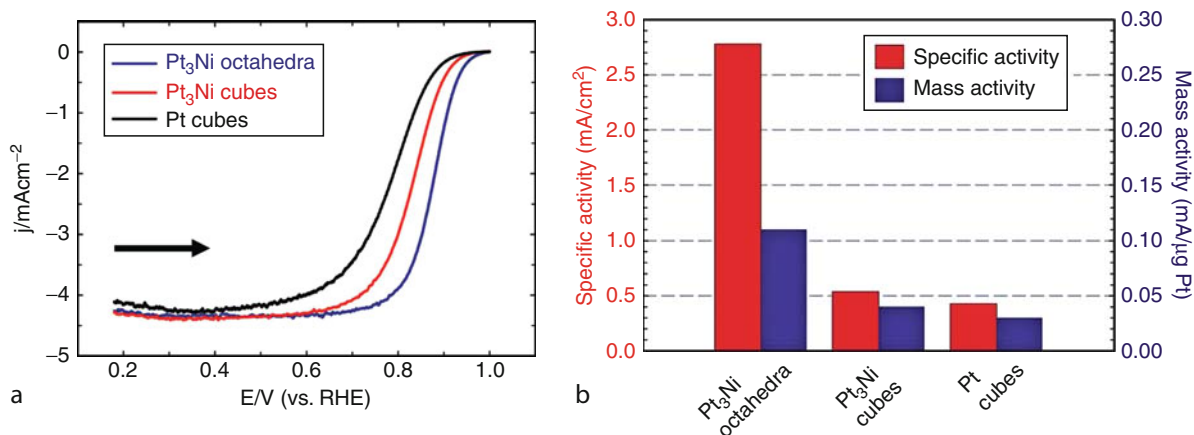
(a–e) Images for Pt<sub>3</sub>Ni nano-octahedra. (f–j) Images for Pt<sub>3</sub>Ni nanocubes. (a, f) Field-emission SEM images. (b, g) High-resolution SEM images. (c) 3D model of an octahedron. (d, i) TEM images. (e, j) High-resolution TEM images of single nanocrystals. (h) 3D model of a cube (Reproduced from [142]. With permission)

compositions of the crystals were analyzed by using combined ICP-MS and EDS techniques (from both TEM and SEM), and the results suggested that the average molar ratio of Pt over Ni was 3:1.

Zhang et al. [142] further investigated ORR activities of the shape controlled nanocrystals by using RDE measurements. The ORR measurements were conducted in an O<sub>2</sub>-saturated 0.1 M HClO<sub>4</sub> solution at 295 K. A characteristic set of polarization curves at 900 rpm for the ORR on Pt<sub>3</sub>Ni nano-octahedra, Pt<sub>3</sub>Ni nanocubes, and Pt nanocubes are displayed in Fig. 17a. After mass transport correction using Koutecký–Levich equation, and normalizing by the Pt surface area and mass, the kinetic activities (specific activity and mass activity) at 0.9 V were plotted in Fig. 17b. The Pt-specific activity of Pt<sub>3</sub>Ni nano-octahedra were determined to be 5.1 times of that of the Pt<sub>3</sub>Ni nanocubes and ~6.5 times that of the Pt nanocubes, while the Pt mass activity of the Pt<sub>3</sub>Ni nano-octahedra was ~2.8 times of that of Pt<sub>3</sub>Ni nanocubes and ~3.6 times of that of Pt nanocubes. The significant shape dependence of ORR activity agreed with the observation from the extended Pt<sub>3</sub>Ni single crystal surfaces, although the

absolute values of specific activities observed on Pt<sub>3</sub>Ni nanocubes and nano-octahedra were about four- to sevenfold lower than those reported in Ref. [42], respectively. One apparent puzzle in these reported results is that the Pt surface area or ECSA of the nanocrystals derived from the specific activity and mass activity (ECSA per unit mass of Pt = mass activity/specific activity) is 5–10 times lower than one would expect from the size of particles revealed by SEM and TEM images. The discrepancy may come from the low utilization of the surface area because of impurities or from overlap of the nonsupported nanocrystals. Another set of data for the high surface carbon supported those nanocrystals was reported in the online supporting information of the paper, and this showed a better agreement between measured ECSA and diameter to TEM, suggesting that particle aggregation caused the low area observed for the unsupported catalysts.

In summary, the size- and shape-controlled synthesis of nanocrystal Pt-based electrocatalysts has shown a promising path to high Pt-specific activity, although the absolute number of the activity is still not comparable



### PEM Fuel Cells and Platinum-Based Electrocatalysts. Figure 17

(a) Polarization curves for ORR on Pt<sub>3</sub>Ni nano-octahedra, Pt<sub>3</sub>Ni nanocubes, and Pt nanocubes supported on a rotating glassy carbon disk electrode in O<sub>2</sub>-saturated 0.1 M HClO<sub>4</sub> solution at 295 K; with scan rate = 20 mV/s; rotation rate = 900 rpm. Catalyst loading in terms of Pt mass: Pt<sub>3</sub>Ni octahedra, 3.0 μg; Pt<sub>3</sub>Ni cube, 2.0 μg; Pt cube, 1.1 μg. Current density was normalized to the glassy carbon geometric surface area (0.196 cm<sup>2</sup>). The arrow indicates the potential scan direction.

(b) Comparison of the ORR activities on the three types of catalysts. Specific activity and mass activity were all measured at 0.9 V versus RHE at 295 K (Reproduced from [142]. With permission)

to that observed on extended Pt alloy single-crystal surface, probably due to the size effect, residual impurities and defects in the nanocrystal surface, and incomplete formation of smooth, segregated Pt layers on the facet surfaces. The Pt mass activity achieved for the best case is already about four times higher than state-of-the-art Pt/C catalyst. If the core of the nanocrystals can be replaced with some corrosion-resistant material but keeping the surface of the Pt alloy shell still in {111} facets, one could expect a significant further reduction of the Pt loading required for the cathode catalyst. In addition, the size of the particles can be larger to gain the benefits of the particle size effect, as the Pt consumption is determined by the thickness of the shell. The durability of such nanocrystals could be expected to be high because of the lack of low-coordination atoms in their surfaces.

### Future Directions

Low platinum loading, high activity, and more durable catalysts still remain as critical challenges for PEFCs for automotive applications. Further fundamental understanding of the correlations between activity, stability, and structural properties at the atomic level are most desired from both theoretical and experimental

perspectives. Studies of the connections between the activities of controlled-facet-orientation nanoparticles and extended single-crystal surfaces would be helpful. Structure- and surface-controlled syntheses of catalysts (Pt monolayer catalysts, nanostructured catalysts and electrodes, size- and facet- controlled Pt alloy nanocrystals, combined with core-shell structure) should provide a practical viable path to achieving fuel cell catalyst loadings required for large-scale commercialization.

### Bibliography

#### Primary Literature

1. Andujar JM, Segura F (2009) Fuel cells: history and updating. A walk along two centuries. *Renew Sustain Energy Rev* 13(9):2309–2322
2. Grimes PG (2000) Historical pathways for fuel cells – the new electric century. *IEEE Aerosp Electron Syst Mag* 15(12):7–10
3. Appleby AJ (1990) From Sir William Grove to today: fuel cells and the future. *J Power Sources* 29(1–2):3–11
4. Perry ML, Fuller TF (2002) A historical perspective of fuel cell technology in the 20th century. *J Electrochem Soc* 149(7): S59–S67
5. Thomas CE (2009) Fuel cell and battery electric vehicles compared. *Int J Hydrogen Energy* 34(15):6005–6020

- Gottesfeld S (2007) Fuel cell techno-personal milestones 1984–2006. *J Power Sources* 171(1):37–45
- Mathias MF, Makharia R, Gasteiger HA, Conley JJ, Fuller TJ, Gittleman CJ, Kocha SS, Miller DP, Mittelsteadt CK, Xie T, Yan SG, Yu PT (2005) Two fuel cell cars in every garage? *Electrochem Soc Interface* 14(3 Fall 2005):24–35
- Raistrick ID (1986) In: Zee JWV, White RE, Kinoshita K, Burney HS (eds) Diaphragms, separators, and ion-exchange membranes, the electrochemical society proceedings series. The Electrochemical Society, Pennington, p 172
- Wilson MS, Gottesfeld S (1992) Thin-film catalyst layers for polymer electrolyte fuel cell electrodes. *J Appl Electrochem* 22(1):1–7
- Wilson MS, Gottesfeld S (1992) High performance catalyzed membranes of ultra-low Pt loadings for polymer electrolyte fuel cells. *J Electrochem Soc* 139(2):L28–L30
- Wilson MS, Valerio JA, Gottesfeld S (1995) Low platinum loading electrodes for polymer electrolyte fuel-cells fabricated using thermoplastic ionomers. *Electrochim Acta* 40(3):355–363
- Conway BE, Tilak BV (2002) Interfacial processes involving electrocatalytic evolution and oxidation of H<sub>2</sub>, and the role of chemisorbed H. *Electrochim Acta* 47(22–23):3571–3594
- Gasteiger HA, Markovic NM, Ross PN (1995) H<sub>2</sub> and CO electrooxidation on well-characterized Pt, Ru, and Pt-Ru. 2. rotating disk electrode studies of CO/H<sub>2</sub> mixtures at 62-degrees C. *J Phys Chem* 99(45):16757–16767
- Mukerjee S, McBreen J (1996) Hydrogen electrocatalysis by carbon supported Pt and Pt alloys – an in situ x-ray absorption study. *J Electrochem Soc* 143(7):2285–2294
- Neyerlin KC, Gu WB, Jorne J, Gasteiger HA (2007) Study of the exchange current density for the hydrogen oxidation and evolution reactions. *J Electrochem Soc* 154(7):B631–B635
- Tarasevich MR, Sadkowski A, Yeager E (1983) Oxygen electrochemistry. In: Conway BE, Bockris JO, Yeager E, Khan SUM, White RE (eds) Comprehensive treatise in electrochemistry. Plenum Press, New York, p 301
- Adzic RR (1998) Recent advances in the kinetics of oxygen reduction. In: Lipkowsky J, Ross PN (eds) *Electrocatalysis*. Wiley-VCH, New York, pp 197–241
- Kinoshita K (1992) *Electrochemical oxygen technology*. Wiley, New York
- Markovic NM, Gasteiger HA, Ross PN (1995) Oxygen reduction on platinum low-index single-crystal surfaces in sulfuric-acid-solution – rotating ring-Pt(Hkl) disk studies. *J Phys Chem* 99(11):3411–3415
- Gasteiger HA, Panels JE, Yan SG (2004) Dependence of PEM fuel cell performance on catalyst loading. *J Power Sources* 127(1–2):162–171
- Gasteiger HA, Kocha SS, Sompalli B, Wagner FT (2005) Activity benchmarks and requirements for Pt, Pt-alloy, and non-Pt oxygen reduction catalysts for PEMFCs. *Appl Catal B Environ* 56(1–2):9–35
- Damjanovic A, Brusic V (1967) Electrode kinetics of oxygen reduction on oxide-free platinum electrodes. *Electrochim Acta* 12(6):615–628
- Wang JX, Markovic NM, Adzic RR (2004) Kinetic analysis of oxygen reduction on Pt(111) in acid solutions: intrinsic kinetic parameters and anion adsorption effects. *J Phys Chem B* 108(13):4127–4133
- Markovic NM, Gasteiger HA, Grgur BN, Ross PN (1999) Oxygen reduction reaction on Pt(111): effects of bromide. *J Electroanal Chem* 467(1):157–163
- Adzic RR (1992) Surface morphology effects in oxygen electrochemistry. In: Scherson D, Tryk D, Xing X (eds) *Proceedings of the workshop on structural effects in electrocatalysis and oxygen electrochemistry*. The Electrochemical Society, Pennington, p 419
- Uribe FA, Wilson MS, Springer TE, Gottesfeld S (1992) Oxygen reduction (ORR) at the Pt/recast ionomer interface and some general comments on the ORR at Pt/aqueous electrolyte interfaces. In: Scherson DD, Tryk D, Xing X (eds) *Proceedings of the workshop on structural effects in electrocatalysis and oxygen electrochemistry*. The Electrochemical Society, Pennington, p 494
- Norskov JK, Rossmeisl J, Logadottir A, Lindqvist L, Kitchin JR, Bligaard T, Jonsson H (2004) Origin of the overpotential for oxygen reduction at a fuel-cell cathode. *J Phys Chem B* 108(46):17886–17892
- Wang JX, Zhang JL, Adzic RR (2007) Double-trap kinetic equation for the oxygen reduction reaction on Pt(111) in acidic media. *J Phys Chem A* 111(49):12702–12710
- Wang JX, Uribe FA, Springer TE, Zhang JL, Adzic RR (2008) Intrinsic kinetic equation for oxygen reduction reaction in acidic media: the double Tafel slope and fuel cell applications. *Faraday Discuss* 140:347–362
- Neyerlin KC, Gu WB, Jorne J, Gasteiger HA (2006) Determination of catalyst unique parameters for the oxygen reduction reaction in a PEMFC. *J Electrochem Soc* 153(10):A1955–A1963
- Neyerlin KC, Gu W, Jorne J, Clark A, Gasteiger HA (2007) Cathode catalyst utilization for the ORR in a PEMFC – analytical model and experimental validation. *J Electrochem Soc* 154(2):B279–B287
- Neyerlin KC, Gasteiger HA, Mittelsteadt CK, Jorne J, Gu WB (2005) Effect of relative humidity on oxygen reduction kinetics in a PEMFC. *J Electrochem Soc* 152(6):A1073–A1080
- Blurton KF, Greenberg P, Oswin HG, Rutt DR (1972) The electrochemical activity of dispersed platinum. *J Electrochem Soc* 119(5):559–564
- Peuckert M, Yoneda T, Betta RAD, Boudart M (1986) Oxygen reduction on small supported platinum particles. *J Electrochem Soc* 133(5):944–947
- Kinoshita K (1990) Particle size effects for oxygen reduction on highly dispersed platinum in acid electrolytes. *J Electrochem Soc* 137(3):845–848
- Ross PN (1986) Structure-property relations in noble metal electrocatalysis. In: *The Gordon conference on chemistry at interfaces*, Lawrence Berkeley Laboratory, Berkeley/Meriden, July 21–25, 1986. p. LBL-21733
- Ross PN (September 1980) Oxygen reduction on supported Pt alloys and intermetallic compounds in phosphoric acid, final

- report prepared for the electric power research institute. Electric Power Research Institute, Palo Alto, September 1980
38. Sattler ML, Ross PN (1986) The surface structure of Pt crystallites supported on carbon black. *Ultramicroscopy* 20:21–28
  39. Landsman DA, Luczak FJ (2003) Catalyst studies and coating technologies. In: Vielstich W, Gasteiger H, Lamm A (eds) *Handbook of fuel cells*. Wiley, Chichester, p 811
  40. Thompsett D (2003) Pt alloys as oxygen reduction catalysts. In: Vielstich W, Gasteiger H, Lamm A (eds) *Handbook of fuel cells – fundamentals, technology and applications*. Wiley, Chichester, p 467
  41. Markovic N, Gasteiger H, Ross PN (1997) Kinetics of oxygen reduction on Pt(hkl) electrodes: implications for the crystallite size effect with supported Pt electrocatalysts. *J Electrochem Soc* 144(5):1591–1597
  42. Stamenkovic VR, Fowler B, Mun BS, Wang GF, Ross PN, Lucas CA, Markovic NM (2007) Improved oxygen reduction activity on Pt<sub>3</sub>Ni(111) via increased surface site availability. *Science* 315(5811):493–497
  43. Hammer B, Norskov JK (2000) Theoretical surface science and catalysis – calculations and concepts. In: Gates BC, Knozinger H (eds) *Advances in catalysis*, vol 45. Academic, San Diego, pp 71–129
  44. Norskov JK, Bligaard T, Logadottir A, Bahn S, Hansen LB, Bollinger M, Bengaard H, Hammer B, Slijivanicanin Z, Mavrikakis M, Xu Y, Dahl S, Jacobsen CJH (2002) Universality in heterogeneous catalysis. *J Catal* 209(2):275–278
  45. Lopez N, Janssens TVW, Clausen BS, Xu Y, Mavrikakis M, Bligaard T, Norskov JK (2004) On the origin of the catalytic activity of gold nanoparticles for low-temperature CO oxidation. *J Catal* 223(1):232–235
  46. Xu Y, Mavrikakis M (2003) Adsorption and dissociation of O<sub>2</sub> on gold surfaces: effect of steps and strain. *J Phys Chem B* 107(35):9298–9307
  47. Xu Y, Ruban AV, Mavrikakis M (2004) Adsorption and dissociation of O<sub>2</sub> on Pt-Co and Pt-Fe alloys. *J Am Chem Soc* 126(14):4717–4725
  48. Greeley J, Rossmeisl J, Hellman A, Norskov JK (2007) Theoretical trends in particle size effects for the oxygen reduction reaction. *Z Phys Chemie-Int J Res Phys Chem Chem Phys* 221(9–10):1209–1220
  49. Mukerjee S, McBreen J (1998) Effect of particle size on the electrocatalysis by carbon-supported Pt electrocatalysts: an in situ XAS investigation. *J Electroanal Chem* 448(2):163–171
  50. Yano H, Inukai J, Uchida H, Watanabe M, Babu PK, Kobayashi T, Chung JH, Oldfield E, Wieckowski A (2006) Particle-size effect of nanoscale platinum catalysts in oxygen reduction reaction: an electrochemical and Pt-195 EC-NMR study. *Phys Chem Chem Phys* 8(42):4932–4939
  51. Gasteiger HA, Markovic NM (2009) Just a dream-or future reality? *Science* 324(5923):48–49
  52. Mukerjee S, Srinivasan S, Soriaga MP, McBreen J (1995) Role of structural and electronic-properties of Pt and Pt alloys on electrocatalysis of oxygen reduction – an in-situ Xanes and EXAFS investigation. *J Electrochem Soc* 142(5):1409–1422
  53. Wakabayashi N, Takeichi M, Uchida H, Watanabe M (2005) Temperature dependence of oxygen reduction activity at Pt-Fe, Pt-Co, and Pt-Ni alloy electrodes. *J Phys Chem B* 109(12):5836–5841
  54. Paulus UA, Wokaun A, Scherer GG, Schmidt TJ, Stamenkovic V, Markovic NM, Ross PN (2002) Oxygen reduction on high surface area Pt-based alloy catalysts in comparison to well defined smooth bulk alloy electrodes. *Electrochim Acta* 47(22–23):3787–3798
  55. Glass JT, Cahen JGL, Stoner GE, Taylor EJ (1987) The effect of metallurgical variables on the electrocatalytic properties of PtCr alloys. *J Electrochem Soc* 134(1):58–65
  56. Paffett MT, Daube KA, Gottesfeld S, Campbell CT (1987) Electrochemical and surface science investigations of PtCr alloy electrodes. *J Electroanal Chem* 220(2):269–285
  57. Beard BC, Ross JPN (1990) The structure and activity of Pt-Co alloys as oxygen reduction electrocatalysts. *J Electrochem Soc* 137(11):3368–3374
  58. Toda T, Igarashi H, Uchida H, Watanabe M (1999) Enhancement of the electroreduction of oxygen on Pt alloys with Fe, Ni, and Co. *J Electrochem Soc* 146(10):3750–3756
  59. Koh S, Hahn N, Yu CF, Strasser P (2008) Effects of composition and annealing conditions on catalytic activities of dealloyed Pt-Cu nanoparticle electrocatalysts for PEMFC. *J Electrochem Soc* 155(12):B1281–B1288
  60. Schulenburg H, Muller E, Khelashvili G, Roser T, Bonnemann H, Wokaun A, Scherer GG (2009) Heat-treated PtCo<sub>3</sub> nanoparticles as oxygen reduction catalysts. *J Phys Chem C* 113(10):4069–4077
  61. Jalan V, Taylor EJ (1983) Importance of interatomic spacing in catalytic reduction of oxygen in phosphoric acid. *J Electrochem Soc* 130(11):2299–2302
  62. Jalan V, Taylor EJ (1984) Importance of interatomic spacing in the catalytic reduction of oxygen in phosphoric acid. In: McIntyre JDE, Weaver MJ, Yeager EB (eds) *The Electrochemical Society Softbound Proceedings Series*. The Electrochemical Society, Pennington, p 546
  63. Landsman DA, Luczak FJ (1982) Noble metal-chromium alloy catalysts and electrochemical cell. US Patent 4,316,944, United Technologies Corporation: US
  64. Stamenkovic VR, Mun BS, Arenz M, Mayrhofer KJJ, Lucas CA, Wang GF, Ross PN, Markovic NM (2007) Trends in electrocatalysis on extended and nanoscale Pt-bimetallic alloy surfaces. *Nat Mater* 6(3):241–247
  65. Toda T, Igarashi H, Watanabe M (1999) Enhancement of the electrocatalytic O<sub>2</sub> reduction on Pt-Fe alloys. *J Electroanal Chem* 460(1–2):258–262
  66. M-k M, Cho J, Cho K, Kim H (2000) Particle size and alloying effects of Pt-based alloy catalysts for fuel cell applications. *Electrochim Acta* 45(25–26):4211–4217
  67. Koh S, Strasser P (2007) Electrocatalysis on bimetallic surfaces: modifying catalytic reactivity for oxygen reduction by voltammetric surface dealloying. *J Am Chem Soc* 129(42):12624

68. Gottesfeld S (1986) The ellipsometric characterization of Pt + Cr alloy surfaces in acid solutions. *J Electroanal Chem* 205(1–2):163–184
69. Paffett MT, Beery JG, Gottesfeld S (1988) Oxygen reduction at Pt<sub>0.65</sub>Cr<sub>0.35</sub>, Pt<sub>0.2</sub>Cr<sub>0.8</sub> and roughened platinum. *J Electrochem Soc* 135(6):1431–1436
70. Mukerjee S, Srinivasan S (1993) Enhanced electrocatalysis of oxygen reduction on platinum alloys in proton exchange membrane fuel cells. *J Electroanal Chem* 357(1–2): 201–224
71. Toda T, Igarashi H, Watanabe M (1998) Role of electronic property of Pt and Pt alloys on electrocatalytic reduction of oxygen. *J Electrochem Soc* 145(12):4185–4188
72. Mun BS, Watanabe M, Rossi M, Stamenkovic V, Markovic NM, Ross PN (2005) A study of electronic structures of Pt3M (M = Ti, V, Cr, Fe, Co, Ni) polycrystalline alloys with valence-band photoemission spectroscopy. *J Chem Phys* 123(20):204717
73. Greeley J, Stephens IEL, Bondarenko AS, Johansson TP, Hansen HA, Jaramillo TF, Rossmeisl J, Chorkendorff I, Norskov JK (2009) Alloys of platinum and early transition metals as oxygen reduction electrocatalysts. *Nat Chem* 1(7):552–556
74. Mukerjee S, Srinivasan S, Soriaga MP, McBreen J (1995) Effect of preparation conditions of Pt Alloys on their electronic, structural, and electrocatalytic activities for oxygen reduction-XRD, XAS, and electrochemical studies. *J Phys Chem* 99(13):4577–4589
75. Uribe FA, Zawodzinski TA (2002) A study of polymer electrolyte fuel cell performance at high voltages. Dependence on cathode catalyst layer composition and on voltage conditioning. *Electrochim Acta* 47(22–23):3799–3806
76. Stamenkovic V, Schmidt TJ, Ross PN, Markovic NM (2002) Surface composition effects in electrocatalysis: kinetics of oxygen reduction on well-defined Pt3Ni and Pt3Co alloy surfaces. *J Phys Chem B* 106(46):11970–11979
77. Murthi VS, Urian RC, Mukerjee S (2004) Oxygen reduction kinetics in low and medium temperature acid environment: Correlation of water activation and surface properties in supported Pt and Pt alloy electrocatalysts. *J Phys Chem B* 108(30):11011–11023
78. Teliska M, Murthi VS, Mukerjee S, Ramaker DE (2005) Correlation of water activation, surface properties, and oxygen reduction reactivity of supported Pt-M/C bimetallic electrocatalysts using XAS. *J Electrochem Soc* 152(11):A2159–A2169
79. Lima FHB, Ticianelli EA (2004) Oxygen electrocatalysis on ultrathin porous coating rotating ring/disk platinum and platinum-cobalt electrodes in alkaline media. *Electrochim Acta* 49(24):4091–4099
80. Lima FHB, Giz MJ, Ticianelli EA (2005) Electrochemical performance of dispersed Pt-M (M = V, Cr and Co) nanoparticles for the oxygen reduction electrocatalysis. *J Braz Chem Soc* 16(3 A):328–336
81. Lima FHB, Salgado JRC, Gonzalez ER, Ticianelli EA (2007) Electrocatalytic properties of PtCoC and PtNiC alloys for the oxygen reduction reaction in alkaline solution. *J Electrochem. So.* 154(4)
82. Creemers C, Deurinck P (1997) Platinum segregation to the (111) surface of ordered Pt<sub>80</sub>Fe<sub>20</sub>: LEIS results and model simulations. *Surf Interface Anal* 25(3):177–189
83. Gauthier Y, Joly Y, Baudoing R, Rundgren J (1985) Surface-sandwich segregation on nondilute bimetallic alloys: Pt50Ni50 and Pt78Ni22 probed by low-energy electron diffraction. *Phys Rev B* 31(10):6216–6218
84. Gauthier Y, Baudoing-Savois R, Bugnard JM, Hebenstreit W, Schmid M, Varga P (2000) Segregation and chemical ordering in the surface layers of Pt<sub>25</sub>Co<sub>75</sub>(111): a LEED/STM study. *Surf Sci* 466(1–3):155–166
85. Gasteiger HA, Ross PN Jr, Cairns EJ (1993) LEIS and AES on sputtered and annealed polycrystalline Pt-Ru bulk alloys. *Surf Sci* 293(1–2):67–80
86. Ruban AV, Skriver HL, Norskov JK (1999) Surface segregation energies in transition-metal alloys. *Phys Rev B* 59(24):15990–16000
87. Ma Y, Balbuena PB (2008) Pt surface segregation in bimetallic Pt3M alloys: a density functional theory study. *Surf Sci* 602(1):107–113
88. Chen S, Ferreira PJ, Sheng WC, Yabuuchi N, Allard LF, Shao-Horn Y (2008) Enhanced activity for oxygen reduction reaction on “Pt3Co” nanoparticles: direct evidence of percolated and sandwich-segregation structures. *J Am Chem Soc* 130(42):13818–13819
89. Stamenkovic VR, Mun BS, Mayrhofer KJJ, Ross PN, Markovic NM (2006) Effect of surface composition on electronic structure, stability, and electrocatalytic properties of Pt-transition metal alloys: Pt-skin versus Pt-skeleton surfaces. *J Am Chem Soc* 128(27):8813–8819
90. Chen S, Sheng WC, Yabuuchi N, Ferreira PJ, Allard LF, Shao-Horn Y (2009) Origin of oxygen reduction reaction activity on “Pt3Co” nanoparticles: atomically resolved chemical compositions and structures. *J Phys Chem C* 113(3): 1109–1125
91. Koh S, Leisch J, Toney MF, Strasser P (2007) Structure-activity-stability relationships of Pt-Co alloy electrocatalysts in gas-diffusion electrode layers. *J Phys Chem C* 111(9): 3744–3752
92. Mani P, Srivastava R, Strasser P (2008) Dealloyed Pt-Cu core-shell nanoparticle electrocatalysts for use in PEM fuel cell cathodes. *J Phys Chem C* 112(7):2770–2778
93. Srivastava R, Mani P, Hahn N, Strasser P (2007) Efficient oxygen reduction fuel cell electrocatalysis on voltammetrically dealloyed Pt-Cu-Co nanoparticles. *Angew Chem Int Ed Engl* 46(47):8988–8991
94. Neyerlin KC, Srivastava R, Yu CF, Strasser P (2009) Electrochemical activity and stability of dealloyed Pt-Cu and Pt-Cu-Co electrocatalysts for the oxygen reduction reaction (ORR). *J Power Sources* 186(2):261–267
95. Wang C, Van Der Vliet D, Chang KC, You H, Strmcnik D, Schlueter JA, Markovic NM, Stamenkovic VR (2009) Monodisperse Pt3Co nanoparticles as a catalyst for the oxygen reduction reaction: size-dependent activity. *J Phys Chem C* 113(45):19365–19368



96. Watanabe M, Wakisaka M, Yano H, Uchida H (2008) Analyses of oxygen reduction reaction at Pt-based electrocatalysts. *ECS Trans* 16:199–206
97. Wakisaka M, Suzuki H, Mitsui S, Uchida H, Watanabe M (2008) Increased oxygen coverage at Pt-Fe alloy cathode for the enhanced oxygen reduction reaction studied by EC-XPS. *J Phys Chem C* 112(7):2750–2755
98. Ferreira PJ, Ia O GJ, Shao-Horn Y, Morgan D, Makharia R, Kocha S, Gasteiger HA (2005) Instability of Pt/C electrocatalysts in proton exchange membrane fuel cells – a mechanistic investigation. *J Electrochem Soc* 152(11):A2256–A2271
99. Colon-Mercado HR, Popov BN (2006) Stability of platinum based alloy cathode catalysts in PEM fuel cells. *J Power Sources* 155(2):253–263
100. Morita T, Kojima K (2008) Development of fuel cell hybrid vehicle in Toyota. *ECS Trans* 16:185–198
101. Uchimura M, Sugawara S, Suzuki Y, Zhang J, Kocha SS (2008) Electrocatalyst durability under simulated automotive drive cycles. *ECS Trans* 16(2):225–234
102. Adzic RR, Zhang J, Sasaki K, Vukmirovic MB, Shao M, Wang JX, Nilekar AU, Mavrikakis M, Valerio JA, Uribe F (2007) Platinum monolayer fuel cell electrocatalysts. *Top Catal* 46(3–4):249–262
103. Brankovic SR, Wang JX, Adzic RR (2001) Pt submonolayers on Ru nanoparticles – a novel low Pt loading, high CO tolerance fuel cell electrocatalyst. *Electrochem Solid State Lett* 4(12):A217–A220
104. Sasaki K, Mo Y, Wang JX, Balasubramanian M, Uribe F, McBreen J, Adzic RR (2003) Pt submonolayers on metal nanoparticles – novel electrocatalysts for H<sub>2</sub> oxidation and O<sub>2</sub> reduction. *Electrochim Acta* 48(25–26):3841–3849
105. Wang JX, Brankovic SR, Zhu Y, Hanson JC, Adzic RR (2003) Kinetic characterization of PtRu fuel cell anode catalysts made by spontaneous Pt deposition on Ru nanoparticles. *J Electrochem Soc* 150(8):A1108–A1117
106. Brankovic SR, McBreen J, Adzic RR (2001) Spontaneous deposition of Pt on the Ru(0001) surface. *J Electroanal Chem* 503(1–2):99–104
107. Sasaki K, Wang JX, Balasubramanian M, McBreen J, Uribe F, Adzic RR (2004) Ultra-low platinum content fuel cell anode electrocatalyst with a long-term performance stability. *Electrochim Acta* 49(22–23 SPEC. ISS):3873–3877
108. Kolb DM, Przasnyski M, Gerischer H (1974) Underpotential deposition of metals and work function differences. *J Electroanal Chem* 54(1):25–38
109. Herrero E, Buller LJ, Abruna HD (2001) Underpotential deposition at single crystal surfaces of Au, Pt, Ag and other materials. *Chem Rev* 101(7):1897–1930
110. Aramata A (1997) Underpotential deposition on single-crystal metals. In: Bockris JO, White RE, Conway BE (eds) *Modern aspects of electrochemistry*. Plenum, New York
111. Brankovic SR, Wang JX, Adzic RR (2001) Metal monolayer deposition by replacement of metal adlayers on electrode surfaces. *Surf Sci* 474(1–3):L173–L179
112. Zhang J, Mo Y, Vukmirovic MB, Klie R, Sasaki K, Adzic RR (2004) Platinum monolayer electrocatalysts for O<sub>2</sub> reduction: Pt monolayer on Pd(111) and on carbon-supported Pd nanoparticles. *J Phys Chem B* 108(30):10955–10964
113. Zhang J, Vukmirovic MB, Sasaki K, Uribe F, Adzic RR (2005) Platinum monolayer electro catalysts for oxygen reduction: effect of substrates, and long-term stability. *J Serb Chem Soc* 70(3):513–525
114. Zhang JL, Vukmirovic MB, Xu Y, Mavrikakis M, Adzic RR (2005) Controlling the catalytic activity of platinum-monolayer electrocatalysts for oxygen reduction with different substrates. *Angew Chem Int Ed Engl* 44(14):2132–2135
115. Zhang JL, Vukmirovic MB, Sasaki K, Nilekar AU, Mavrikakis M, Adzic RR (2005) Mixed-metal Pt monolayer electrocatalysts for enhanced oxygen reduction kinetics. *J Am Chem Soc* 127(36):12480–12481
116. Zhou WP, Yang XF, Vukmirovic MB, Koel BE, Jiao J, Peng GW, Mavrikakis M, Adzic RR (2009) Improving electrocatalysts for O<sub>2</sub> reduction by fine-tuning the Pt-support interaction: Pt monolayer on the surfaces of a Pd<sub>3</sub>Fe(111) single-crystal alloy. *J Am Chem Soc* 131(35):12755–12762
117. Zhang J, Lima FHB, Shao MH, Sasaki K, Wang JX, Hanson J, Adzic RR (2005) Platinum monolayer on nonnoble metal-noble metal core-shell nanoparticle electrocatalysts for O<sub>2</sub> reduction. *J Phys Chem B* 109(48):22701–22704
118. Zhang J, Sasaki K, Sutter E, Adzic RR (2007) Stabilization of platinum oxygen-reduction electrocatalysts using gold clusters. *Science* 315(5809):220–222
119. Wang JX, Inada H, Wu LJ, Zhu YM, Choi YM, Liu P, Zhou WP, Adzic RR (2009) Oxygen reduction on well-defined core-shell nanocatalysts: particle size, facet, and Pt shell thickness effects. *J Am Chem Soc* 131(47):17298–17302
120. Sasaki K, Wang JX, Naohara H, Marinkovic N, More K, Inada H, Adzic RR (2010) Recent advances in platinum monolayer electrocatalysts for oxygen reduction reaction: scale-up synthesis, structure and activity of Pt shells on Pd cores. *Electrochim Acta* 55(8):2645–2652
121. Shao-Horn Y, Sheng WC, Chen S, Ferreira PJ, Holby EF, Morgan D (2007) Instability of supported platinum nanoparticles in low-temperature fuel cells. *Top Catal* 46(3–4):285–305
122. Yu PT, Gu W, Makharia R, Wagner FT, Gasteiger HA (2006) The impact of carbon stability on PEM fuel cell startup and shutdown voltage degradation. *ECS Trans* 3:797–809
123. Yu PT, Kocha S, Paine L, Gu W, Wagner FT (2004) The effects of air purge on the degradation of PEM fuel cells during startup and shutdown procedures. In: 2004 AIChE spring national meeting, conference proceedings, New Orleans, pp 521–527
124. Debe MK (2003) Novel catalyst, catalyst support and catalyst coated membrane methods. In: Vielstich W, Gasteiger HA, Lamm A (eds) *Handbook of fuel cells – fundamentals technology and applications*. Wiley, Chichester
125. Gancs L, Kobayashi T, Debe MK, Atanasoski R, Wieckowski A (2008) Crystallographic characteristics of nanostructured thin-film fuel cell electrocatalysts: a HRTEM study. *Chem Mater* 20(7):2444–2454

126. Debe MK, Drube AR (1995) Structural characteristics of a uniquely nanostructured organic thin film. *J Vac Sci Technol B Microelectron Nanometer Struct* 13(3):1236–1241
127. Debe MK, Schmoeckel AK, Vernstrom GD, Atanasoski R (2006) High voltage stability of nanostructured thin film catalysts for PEM fuel cells. *J Power Sources* 161(2):1002–1011
128. Bonakdarpour A, Stevens K, Vernstrom GD, Atanasoski R, Schmoeckel AK, Debe MK, Dahn JR (2007) Oxygen reduction activity of Pt and Pt-Mn-Co electrocatalysts sputtered on nano-structured thin film support. *Electrochim Acta* 53(2):688–694
129. Debe MK, Schmoeckel AK, Hendricks SM, Vernstrom GD, Haugen GM, Atanasoski RT (2005) Durability aspects of nano-structured thin film catalysts for PEM fuel cells. *ECS Trans* 1:51–66
130. Chen ZW, Waje M, Li WZ, Yan YS (2007) Supportless Pt and PtPd nanotubes as electrocatalysts for oxygen-reduction reactions. *Angew Chem Int Edit Engl* 46(22):4060–4063
131. Mayers B, Jiang X, Sunderland D, Cattle B, Xia Y (2003) Hollow nanostructures of platinum with controllable dimensions can be synthesized by templating against selenium nanowires and colloids. *J Am Chem Soc* 125(44):13364–13365
132. Sun Y, Tao Z, Chen J, Herricks T, Xia Y (2004) Ag nanowires coated with Ag/Pd alloy sheaths and their use as substrates for reversible absorption and desorption of hydrogen. *J Am Chem Soc* 126(19):5940–5941
133. Sun Y, Yin Y, Mayers BT, Herricks T, Xia Y (2002) Uniform silver nanowires synthesis by reducing AgNO<sub>3</sub> with ethylene glycol in the presence of seeds and poly(vinyl pyrrolidone). *Chem Mater* 14(11):4736–4745
134. Sun SH, Zhang GX, Geng DS, Chen YG, Banis MN, Li RY, Cai M, Sun XL (2010) Direct growth of single-crystal Pt nanowires on Sn@CNT nanocable: 3D electrodes for highly active electrocatalysts. *Chem Eur J* 16(3):829–835
135. Peng ZM, Yang H (2009) Synthesis and oxygen reduction electrocatalytic property of Pt-on-Pd bimetallic heteronanostructures. *J Am Chem Soc* 131(22):7542
136. Lim B, Jiang M, Camargo PHC, Cho EC, Tao J, Lu X, Zhu Y, Xia Y (2009) Pd-Pt bimetallic nanodendrites with high activity for oxygen reduction. *Science* 324(5932):1302–1305
137. Lim BW, Lu XM, Jiang MJ, Camargo PHC, Cho EC, Lee EP, Xia YN (2008) Facile synthesis of highly faceted multioctahedral Pt nanocrystals through controlled overgrowth. *Nano Lett* 8(11):4043–4047
138. Erlebacher J, Snyder J (2009) Dealloyed nanoporous metals for PEM fuel cell catalysis. *ECS Trans* 25:603–612
139. Zeis R, Mathur A, Fritz G, Lee J, Erlebacher J (2007) Platinum-plated nanoporous gold: An efficient, low Pt loading electrocatalyst for PEM fuel cells. *J Power Sources* 165(1):65–72
140. Erlebacher J (2009) Materials science of hydrogen/oxygen fuel cell catalysis. In: Ehrenreich H, Spaepen F (eds) *Solid state physics – advances in research and applications*. Academic, New York, pp 77–141
141. Wu J, Zhang J, Peng Z, Yang S, Wagner FT, Yang H (2010) Truncated octahedral Pt<sub>3</sub>Ni oxygen reduction reaction electrocatalysts. *J Am Chem Soc* 132(14):4984–4985
142. Zhang J, Yang H, Fang J, Zou S (2010) Synthesis and oxygen reduction activity of shape-controlled Pt<sub>3</sub>Ni nanopolyhedra. *Nano Lett* 10(2):638–644

## Books and Reviews

- Bard AJ, Faulkner LR (2001) *Electrochemical methods, fundamentals and applications*, 2nd edn. Wiley, New York
- Lipkowsky J, Ross P (eds) (1998) *Electrocatalysis (frontiers in electrochemistry)*. Wiley-VCH, Danvers
- Markovic NM, Ross PN Jr (2002) Surface science studies of model fuel cell electrocatalysts. *Surf Sci Rep* 45(4–6):117–229
- Newman J, Thomas-Alyea KE (2004) *Electrochemical system*, 3rd edn. Wiley, Hoboken
- Vielstich W, Gasteiger H, Lamm A (eds) (2003) *Handbook of fuel cells: fundamentals, technology, applications*. Wiley, Chichester
- Vielstich W, Gasteiger H, Lamm A (eds) (2009) *Handbook of fuel cells: advances in electrocatalysis, materials, diagnostics and durability*, vol 5 and 6. Wiley, New York
- Wieckowski A, Savinova ER, Vayenas CG (eds) (2003) *Catalysis and electrocatalysis at nanoparticle surfaces*, 1st edn. CRC Press, Boca Raton
- Zhang J (ed) (2008) *PEM fuel cell electrocatalysts and catalyst layers: fundamentals and applications*, 1st edn. Springer, London

## PEM Fuel Cells, Materials and Design Development Challenges

STEPHEN J. PADDISON<sup>1</sup>, HUBERT A. GASTEIGER<sup>2</sup>

<sup>1</sup>Department of Chemical & Biomolecular Engineering, University of Tennessee, Knoxville, TN, USA

<sup>2</sup>Department of Chemistry, Technische Universität München, Munich, Germany

## Article Outline

Glossary

Definition of the Subject and Its Importance

Introduction

Processes in an MEA and Voltage-Loss Terms

Ion and Water Transport in Ionomers

Degradation of Pt-Based Catalysts

Carbon-Support Corrosion  
 Membrane Development Needs and Approaches  
 Temperature Targets  
 Proton Conductivity  
 Reactant Gas Permeability  
 Morphology  
 Choice of the Protogenic Group  
 Future Directions  
 Bibliography

## Glossary

**Hydrogen oxidation reaction (HOR)** The electrochemical oxidation of molecular hydrogen occurring at the anode of a fuel cell.

**Membrane electrode assembly (MEA)** The assembly consisting of the electrolyte membrane sandwiched between the anode and cathode.

**Oxygen reduction reaction (ORR)** The electrochemical reduction of molecular oxygen through a four electron transfer at the cathode of a fuel cell.

**Perfluorosulfonic acid (PFSA)** The  $\text{CF}_2\text{SO}_3\text{H}$  group is the protogenic group on ionomers and membranes utilized in catalyst layer and electrolyte in a fuel cell.

**Proton exchange membrane (PEM)** A solid polymer thin film that is proton conducting and functions as the central component of a fuel cell.

## Definition of the Subject and Its Importance

Substantial resources have been devoted over the past decade to the development of proton exchange membrane (PEM) fuel cells that use hydrogen fuel and oxygen from the air to produce electricity for applications including automotive propulsion. Remaining challenges include the design of inexpensive and stable robust catalysts for the electrochemical reaction at the cathode (i.e., the reduction of oxygen) of the fuel cell and the synthesis of robust (i.e., chemical and mechanical stable) electrolyte membranes exhibiting high proton conductivity under hot and dry conditions.

## Introduction

The development of commercially viable proton exchange membrane (PEM) fuel cell systems powered by hydrogen or hydrogen-rich reformat faces

a significant number of materials and MEA (membrane electrode assembly) design-related performance and durability challenges, which need to be addressed via:

1. Improvement of current platinum-based catalysts for the oxygen reduction reaction (ORR) and the hydrogen oxidation reaction (HOR) or substitution by platinum-group metal (PGM) free catalysts in order to meet the platinum cost and design constraints for commercial applications [1].
2. Development of more durable ORR and HOR catalysts, which are resistant to the voltage-cycles occurring during the transient operation of fuel cell vehicles (owing to the variable power demand during typical vehicle drive cycles) [2, 3].
3. Substitution of currently used perfluorosulfonic acid (PFSA) ionomers and ionomer membranes (e.g., Nafion<sup>®</sup>) by novel materials with substantially improved proton conductivity at low relative humidity (RH), which would eliminate the need for fully humidified reactants and thereby significantly simplify fuel cell system design [2, 4, 5].
4. Modification/development of ionomers and ionomeric membranes to obtain enhanced chemical durability under low-RH conditions [6, 7] as well as increased mechanical stability during RH-cycles, both of which are frequently occurring conditions under automotive fuel cell operation [2, 8].
5. Replacement of current high-surface area carbon supports (e.g., Ketjen blacks) with more corrosion-resistant materials (e.g., fully graphitized carbon supports or noncarbon-based supports) in order to minimize the damage caused by local hydrogen starvation [9, 10] and during fuel cell start/stop processes [11, 12], so that more complex system-design based mitigation strategies can be avoided [13].
6. Design of HOR catalysts which have no activity for the ORR [14] and their integration into anode electrodes, which is an alternative approach to mitigating degradation caused by start/stop.
7. Mitigation of possible cell-voltage reversal caused by temporary hydrogen under-supply during fast transients, which can be achieved by incorporation of efficient oxygen evolution catalysts into the anode electrode [15] or by corrosion-resistant

- anode catalyst supports (e.g., whisker electrodes developed by 3M [16]).
8. Optimization of electrode and MEA performance with new electrode materials (catalysts, catalyst supports, and ionomers [2]), particularly for high-current density operation with low platinum loadings.
  9. Design of high-performing gas-diffusion media (DM) and microporous layer (MPL) coatings which are resistant to contamination [17] and aging caused by fuel cell system transients (i.e., extensive voltage-cycling, start/stop [18]).
  10. Development of *ab initio* catalyst models, particularly for the ORR catalysis [19] as well as rigorous MEA performance models [20, 21] and in situ diagnostic methods [22–24] in order to provide effective analytical methods required for the screening and implementation of improved electrode materials and MEA designs.

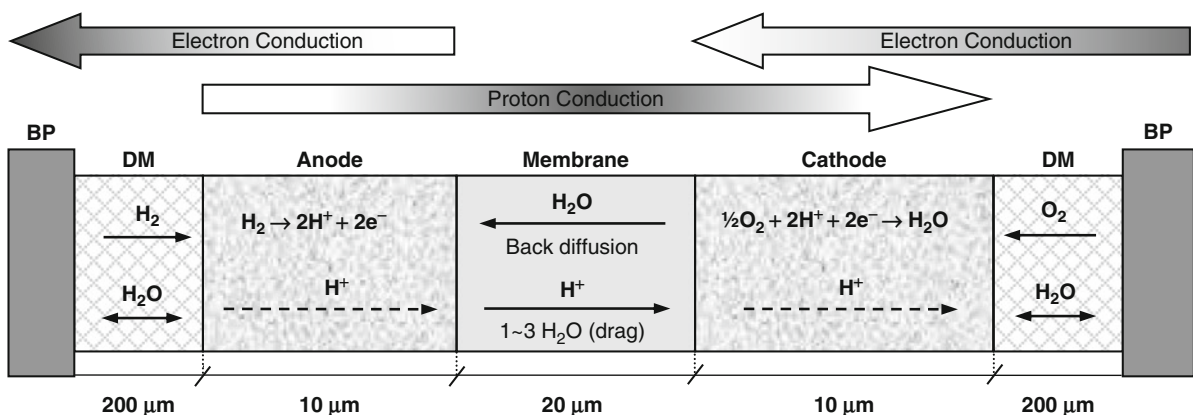
Despite this impressively long list of technical hurdles, the development of fuel cell vehicles has advanced significantly in the past 10 years, and hydrogen fuel cell vehicle fleet tests are now demonstrating drive-cycle efficiencies of 52–58% (lower heating value), real-world vehicle operating life ranging from 700 to 1,900 h, and refueling times on the order of 5 min [25], with the capability of starting from temperatures as low as  $-20^{\circ}\text{C}$  [26]. Nevertheless, in order to meet

the automotive cost targets ( $<50$  US\$ per kW system power), the platinum supply constraints ( $<10$  g platinum per vehicle), and the required durability targets (6,000 operating hours), the above listed challenges remain and the cutting-edge research on these topics is being presented in the subsequent entries of this book.

In the following, we will provide only a very brief overview of some of the basic materials and MEA design concepts and the interested reader is referred to the very detailed articles in the remainder of this encyclopedia or to the cited literature.

### Processes in an MEA and Voltage-Loss Terms

In order to define the limiting factors in fuel cell performance, it is helpful to review the various reactions and transport processes occurring in a PEM fuel cell, which are illustrated in Fig. 1. The HOR and ORR reactions catalyzed by the anode and cathode catalysts, respectively, require facile proton transport through the ionomeric membrane and also throughout the porous electrodes which are typically composed of carbon-supported catalysts and proton-conducting ionomer (exceptions are nanostructured electrodes, for example, those developed by 3M [16]). At the same time, hydrogen and oxygen supplied via flow-field channels in the bipolar plates (BP, see Fig. 1) need to be supplied via gas-phase diffusion through



PEM Fuel Cells, Materials and Design Development Challenges. Figure 1

Schematic of a PEMFC repeating unit showing the electrode reactions,  $\text{H}_2$  and  $\text{O}_2$  gas transport, water transport, as well as proton and electron charge transfer. Typical values of component thickness are shown (not drawn to scale) (Reproduced from W. Gu et al. [20] by permission from Wiley)

the porous gas-diffusion layers [27] and throughout the electrodes. At the design point for automotive operating conditions, that is, at an average humidity of the exiting gas-streams of <100% [20], reactant diffusion can be described by a simple effective diffusion coefficient [28]. However, under conditions where the relative humidity of the exiting gas-streams exceeds 100% (e.g., at fuel cell temperatures below 50°C), the quantitative description of gas transport is more complex due to the presence of liquid water inside the porous layers [29].

The local relative humidity also determines the proton conductivity of the ionomeric membrane [30] and of the electrodes [31], leading to major voltage losses below 50% RH. An additional voltage loss, particularly at high-current densities and low relative humidity, arises from the dry-out of the anode-side of the membrane, which is due to the electro-osmotic drag of water caused by protons flowing from the anode to the cathode (vide infra). Owing to the drastically increasing proton conduction resistance at low RH [30], the associated voltage loss can be substantial, unless membranes are very thin in order to allow for fast water back-transport from the cathode to the anode; for this reason, membranes in state-of-the-art PEMFCs are typically not thicker than 15–20 μm.

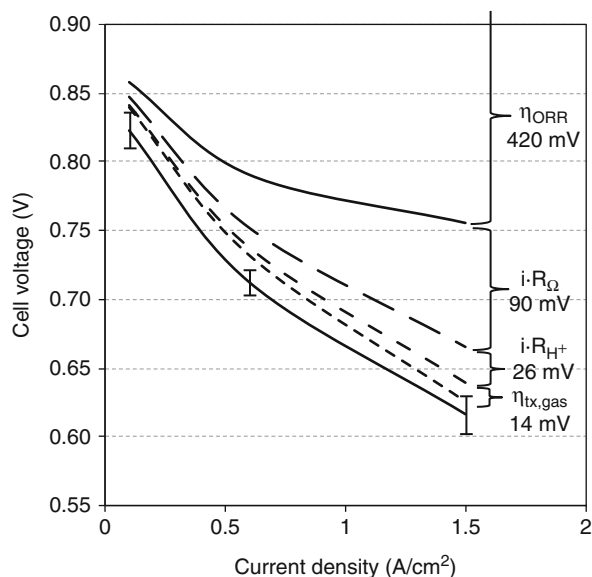
In addition to the voltage losses caused by the overpotentials of the HOR,  $\eta_{HOR}$ , and the ORR,  $\eta_{ORR}$ , as well as by the proton- and gas-transport resistances, substantial voltage losses can arise from purely electronic resistances from the bipolar plate all the way throughout the electrode,  $R_{electronic}$ . However, the bulk resistances of carbon-black based electrodes, carbon-fiber based diffusion media, and stainless steel or carbon-composite based bipolar plates are generally negligible compared to the electronic contact resistances between electrodes/DMs and DMs/BPs [20], so that the measured electronic resistances are mostly due to contact resistances. The largest contribution here comes from the strongly compression-dependent contact resistance between the bipolar plate and carbon-fiber based diffusion media [27].

The above described voltage losses can be summarized by a conceptually simple equation, describing the fuel cell voltage,  $E_{cell}$  as a function of current density,  $i$ :

$$E_{cell} = E_{rev.} - i \times (R_{electronic} + R_{membrane}) - \eta_{HOR} - \eta_{ORR} - i \times R_{H+,eff.} - \eta_{tx,gas(dry)} - \eta_{tx,gas(wet)} \quad (1)$$

where  $E_{rev.}$  is the reversible thermodynamic potential depending on temperature and gas partial pressure. The proton conduction resistances of the membrane,  $R_{membrane}$  and the electrodes,  $R_{H+,eff.}$ , are strongly dependent on the local relative humidity and, in contrast to  $R_{electronic}$ , do also depend on current density and on the local temperature which again is mostly determined by thermal conductivity resistances between the electrode/DM and the DM/bipolar plate interfaces [20, 31, 32]. For fuel cells fed with pure hydrogen and air, the gas-diffusion overpotential,  $\eta_{tx,gas(dry)}$ , is primarily due to the diffusion of oxygen in air through the diffusion medium and the cathode electrode in the absence of liquid water, that is, at operating conditions where the RH of the exiting reactants is <100%; the additional gas-diffusion overpotential losses caused by the presence of liquid water in the diffusion media and the electrodes are here described as  $\eta_{tx,gas(wet)}$  and become significant at wet operating conditions (i.e., at >100% RH of the exiting reactants).

The various voltage losses under typical automotive operating conditions are shown in Fig. 2. At the highest current density of 1.5 A cm<sup>-2</sup>, the voltage loss caused by the slow ORR kinetics amounts to approximately 70% of the overall voltage loss, while the voltage loss for the HOR is negligibly small under these conditions (<<5 mV [33]). At the maximum power density of 0.93 W cm<sup>-2</sup> (0.62 V at 1.5 A cm<sup>-2</sup>) and the total platinum loading of 0.5 g<sub>Pt</sub> cm<sup>-2</sup> shown in Fig. 2, the platinum specific power density is 0.54 g<sub>Pt</sub> kW<sup>-1</sup>, implying that 54 g<sub>Pt</sub> would be required for a typical 100 kW automotive fuel cell. Even though the fast HOR kinetics allow for a lowering of the anode catalyst loading to 0.05 mg<sub>Pt</sub> cm<sup>-2</sup> without notable performance loss [33], the thus obtained platinum specific power density of 0.38 g<sub>Pt</sub> kW<sup>-1</sup> is still too high for commercially viable fuel cells. Therefore, major foci in current fuel cell R&D is the development of either more active platinum-based ORR catalysts or of PGM-free ORR catalysts [1, 34, 35] as well as the development of suitable electrode structures for these novel catalysts. The other voltage loss terms shown in Fig. 2 are



### PEM Fuel Cells, Materials and Design Development Challenges. Figure 2

Voltage loss terms in state-of-the-art  $H_2$ /air PEMFCs operated under representative automotive conditions. MEAs:  $0.2/0.3 \text{ mg}_{Pt} \text{ cm}^{-2}$  (anode/cathode) coated on an  $18 \text{ }\mu\text{m}$  thick composite membrane and sandwiched between  $\approx 200 \text{ }\mu\text{m}$  thick DMs (SGL 25BC). Operating conditions:  $H_2$  and air stoichiometric flows of 2 and 1.8–5.5, respectively, stack pressure of 110–176  $\text{kPa}_{abs}$ , gas inlet humidities of 30–60% RH, and stack temperature of 70–80°C. For details see [20] (Reproduced from W. Gu et al. [20] by permission from Wiley)

significantly smaller and dominated by the ohmic resistances,  $R_{\Omega}$ , which represent  $R_{membrane}$  and  $R_{electronic}$  (s. Eq. 1), whereby 60 of the 90 mV losses at 1.5  $\text{A cm}^{-2}$  are mostly due to the electronic contact resistance between the DMs and the bipolar plates.

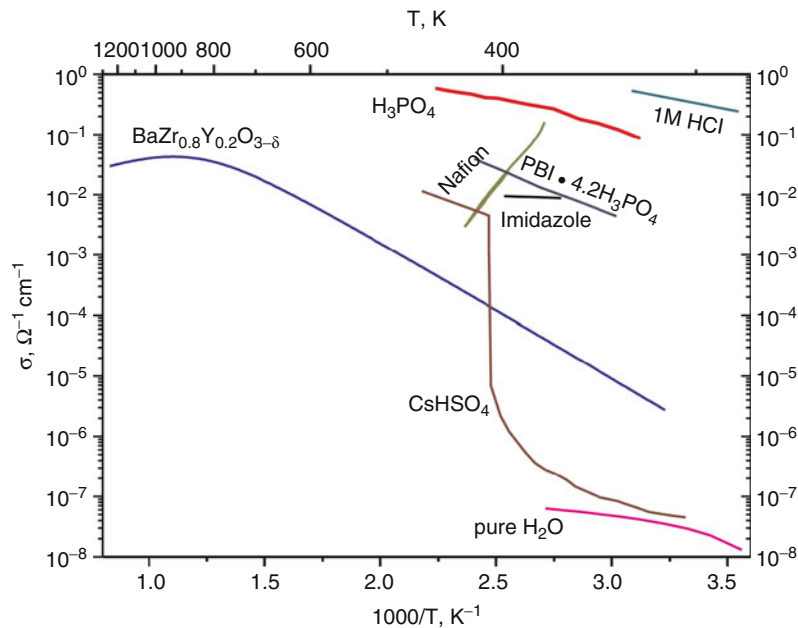
From a systems point of view, it would also be desirable to develop ionomers and ionomeric membranes with increased conductivity at low RH, since this would enable the design of fuel cell systems operating with either dry gas-feeds and/or at higher fuel cell temperature [5]. The latter would reduce the vehicle radiator requirements, which are quite demanding for fuel cell vehicles due to the large heat load which has to be dissipated via the radiator [4], contrary to internal combustion engine-based vehicles where a large fraction of the produced heat is discharged via the latent

heat of the engine exhaust gases. A brief discussion on the membrane development toward this goal is given below.

### Ion and Water Transport in Ionomers

The central component of a PEM fuel cell is a thin film polymer that is an ionomer or ion-containing polymer that critically functions as the separator of gases and electrodes but also as a proton conductor completing the internal circuit. Typically, a proton conductivity of  $0.1 \text{ S cm}^{-1}$  is required for efficient function of the fuel cell. PEMs for automotive applications require that the ionomer function both at high temperature ( $>90^\circ\text{C}$ ) to dissipate waste heat and low pressures ( $<170 \text{ kPa}_{abs}$ ) to minimize pumping parasitics [2]. This necessitates operation at low relative humidities but currently utilized PFSA-based PEMs require water to facilitate the dissociation and long range transport of protons [36]. Although there are a large number of materials and systems that conduct protons (as shown in Fig. 3), those exhibiting sufficient proton conductivity in the target temperature range (for fuel cell operation) are almost nonexistent. It is also interesting to note that while the majority of the materials substances display an increase in conductivity with increasing temperature, the benchmark PFSA ionomer Nafion® exhibits just the opposite if the water vapor pressure remains constant (note: this means that the relative humidity decreases with increasing temperature). This is, of course, due to the dehydration of the membrane as the temperature approaches and then exceeds the boiling point of water (at a water pressure of 1 atm.). All PFSA ionomers show a sharp decrease in proton conductivity as the water content falls (see the entry on “► Proton Exchange Membrane Fuel Cells: High-Temperature, Low-Humidity Operation” by Hamrock and Herring). Several different approaches have been used to improve the proton conductivity of PEMs including changes in the backbone and/or side chain chemistry and are described in detail in the entry “► Membrane Electrolytes, from Perfluoro Sulfonic Acid (PFSA) to Hydrocarbon Ionomers” by Miyatake.

Several approaches have been devised to reduce the resistance of the membrane at low RH, including increasing the density of the acidic groups and making the membrane thinner. There is, of course, a limit to the



PEM Fuel Cells, Materials and Design Development Challenges. Figure 3

Measured proton conductivity for various materials as a function of temperature. In the temperature regime of interest (80–120°C) for PEM fuel cells, the only presently available conductors are the hydrated ionomers (Nafion<sup>®</sup>, etc.) and display a significant decrease in conductivity as the temperature is increased (due to a decrease in water content)

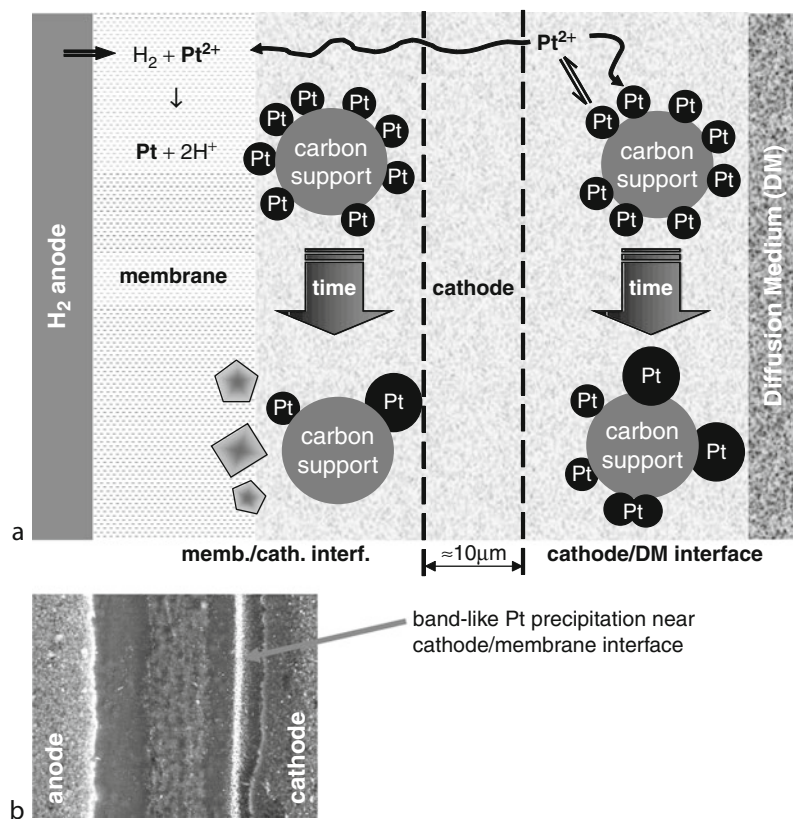
extent to which the acid content may be increased, and the membrane thickness decreased. Membranes with a very high ion exchange capacity suffer mechanical robustness as the low degree of crystallinity and high amounts of water result in materials that dissolve and/or fall apart. Very thin membranes are also prone to mechanical failure and are no longer impermeable to gases which results in both loss of performance and membrane degradation due to the crossover of the reactant gases. Further details of these issues are provided in the entry on “► [High Temperature, Low Humidity Operation of Proton Exchange Membrane Fuel Cells.](#)”

### Degradation of Pt-Based Catalysts

A significant part of the performance degradation of fuel cells with current platinum-based catalysts derives from the dissolution and sintering of platinum. This is due to the relatively high solubility of platinum in the strongly acidic electrolyte [37], whereby the dissolution rate is enhanced by voltage-cycling of the cathode electrode during the dynamic load-following required by

the fuel cell in automotive drive cycles [2, 37]. The enhanced Pt dissolution rate is caused by the transition between mostly metallic platinum at high-current density (viz., at high cathode overpotential) and an oxidized platinum surface at low current density or open circuit potential (viz., at low cathode overpotential) [38, 39]. As illustrated in Fig. 4, dissolved platinum species either redeposit on other platinum particles via an Ostwald ripening process or diffuse in the electrolyte phase toward the membrane, where they precipitate inside the ionomer phase by reaction with hydrogen which is permeating through the membrane from the anode side. Within the ionomer phase, precipitated platinum crystallites form a clearly defined platinum band (see Fig. 4b), the location of which can be predicted by the partial pressure of hydrogen and oxygen in the anode and cathode feed-gases, respectively [40].

Since several hundred thousand large voltage-cycles would be encountered during the lifetime of an automotive fuel cell [2], the associated significant loss of active platinum surface area must be mitigated either by more dissolution-resistant cathode catalysts or by



PEM Fuel Cells, Materials and Design Development Challenges. Figure 4

(a) Schematic representing platinum surface area loss on (i) the nanometer scale, where platinum particles grow on the carbon support via Ostwald ripening and (ii) the micrometer scale, where dissolved platinum species diffuse toward the membrane, become reduced by hydrogen permeating from the anode through the membrane, and precipitate as platinum particles in the membrane. (b) SEM cross section of a short-stack MEA operating at open circuit voltage for 2,000 h, where the bright band in the image indicates platinum deposited in the membrane near the membrane/cathode interface (Reproduced from P. J. Ferreira et al. [37] by permission from The Electrochemical Society)

hybridizing a fuel cell system with a large propulsion battery (tens of kW battery power), by which the number of large voltage-cycles can be substantially reduced. In many instances, it was observed that platinum-alloy catalysts displayed an increased resistance to platinum dissolution [2, 3, 41], but much of the effect is due to the larger particle size of platinum-alloys which favorably affects the platinum dissolution rate via the Gibbs-Thomson effect [42]. Indeed, more recent data demonstrated that voltage-cycling of platinum-alloy cathode catalysts leads to core/shell structures, with platinum-shells forming around a platinum-alloy core, so that the initially higher specific activity of platinum-alloys slowly approaches that of pure

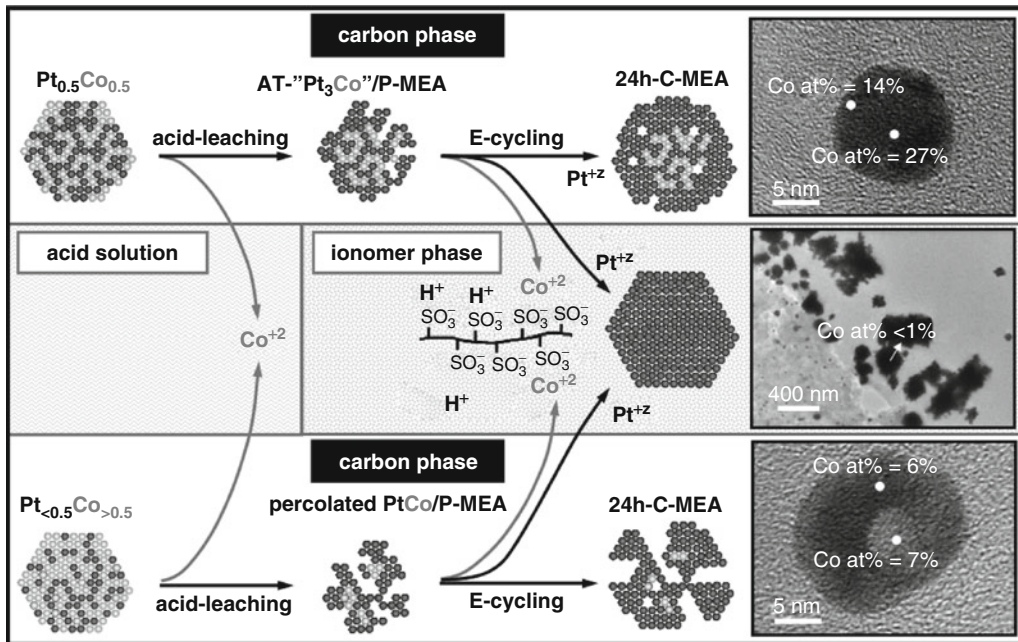
platinum in the course of extensive voltage-cycling [43]. This phenomenon is illustrated in Fig. 5.

In summary, while fuel cell/battery hybrid systems reduce the degradation from platinum surface area loss from voltage-cycling to an acceptable level, novel cathode catalysts with increased stability toward voltage-cycling would bring significant benefits and are therefore a very active field of research.

### Carbon-Support Corrosion

The excellent gas-transport properties of fuel cell electrodes are due to their high porosity, with void volume fractions of  $\approx 60\%$  for the typical ionomer/carbon weight ratios of 1/1 (volume fractions of



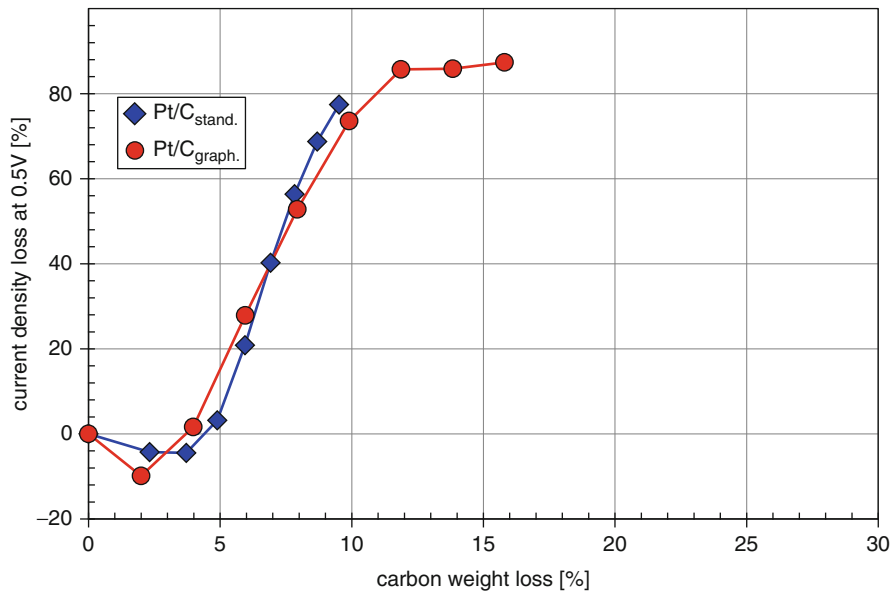


PEM Fuel Cells, Materials and Design Development Challenges. Figure 5

Schematic of the evolution in morphology and composition of a  $\text{Pt}_{0.5}\text{Co}_{0.5}$  cathode catalyst caused by acid leaching and voltage-cycling. The *upper panel* represents particles attached to the carbon-support ("carbon phase"), with *skeleton*  $\text{Pt}_x\text{Co}$  particles obtained after acid leaching, and transforming via Ostwald ripening into *core/shell* particles. The *center panel* represents both the liquid acid phase during pre-leaching and the "ionomer phase" both in the membrane and the electrodes, with large single-crystalline Pt (agglomerates) forming in the membrane and  $\text{Co}^{2+}$  ion-exchanging into the ionomer phase. The *lower panel* is a proposed mechanism for the formation of *percolated*  $\text{Pt}_x\text{Co}$  alloy particles deriving from precursors with higher than average Co content (" $\text{Pt}_{<0.5}\text{Co}_{>0.5}$ ") and resulting in Pt-rich *spongy particles*. TEM bright-field images and spot-resolved EDS compositions (analysis area of 2.5 nm in diameter) of the various types of aged nanoparticles in the 24 h-C-MEA are shown on the right-hand-side (Reproduced from S. Chen et al. [43] by permission from The Electrochemical Society)

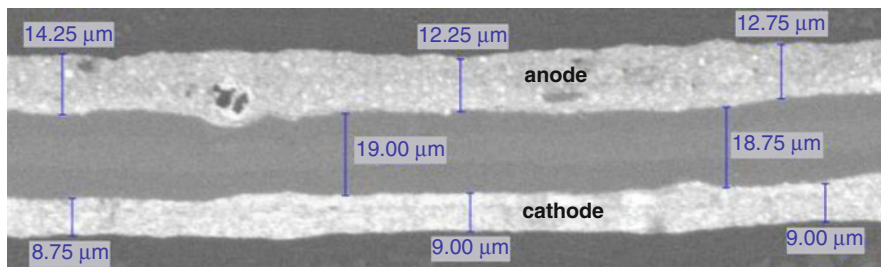
ionomer and carbon being  $\approx 20\%$  each) [20]. This high void volume fraction is due to the use of highly structured carbon blacks (e.g., Vulcan XC72 or Ketjen black), which have an intrinsically low packing density which is maintained up to compressions of  $\approx 100$  MPa (in comparison, MEAs in fuel cell stacks are compressed at  $\approx 1$  MPa). However, at sufficiently high potentials, viz., above  $\approx 1$  V versus the reversible hydrogen electrode (RHE) potential, the electrochemical oxidation of carbon by water ( $\text{C} + 2\text{H}_2\text{O} \rightarrow \text{CO}_2 + 4\text{H}^+ + 4\text{e}^-$ ) is substantial, and after the oxidation of approximately 5–10 wt% of the carbon support, the carbon structure collapses [44], resulting in a rapid decrease of the fuel cell performance, as is shown in Fig. 6.

Under normal fuel cell operating conditions, the highest oxidative potentials in the cathode range between  $\approx 0.6$  V (vs. RHE) at high-current density and  $\approx 0.95$  V (vs. RHE) at open circuit (the anode potential remains always near 0 V vs. RHE), so that carbon-support corrosion is negligible. However, under start/stop conditions or in the case of localized hydrogen starvation, the cathode potential significantly exceeds 1 V versus RHE and the associated rapid carbon-support corrosion leads to a loss of electrode void volume which experimentally is observed as a so-called cathode thinning. This is illustrated by the SEM cross section shown in Fig. 7 for a cathode electrode, where 8 wt% of the carbon support had been oxidized by applying a cathodic potential (the extent



PEM Fuel Cells, Materials and Design Development Challenges. **Figure 6**

Fuel cell performance loss as a function of the extent of carbon-support corrosion. Conditions: H<sub>2</sub>/air ( $s = 2/2$ ) performance at 80°C, 100%RH, 150 kPa<sub>abs</sub> (Reproduced from H.A. Gasteiger et al. [3] by permission from Springer)



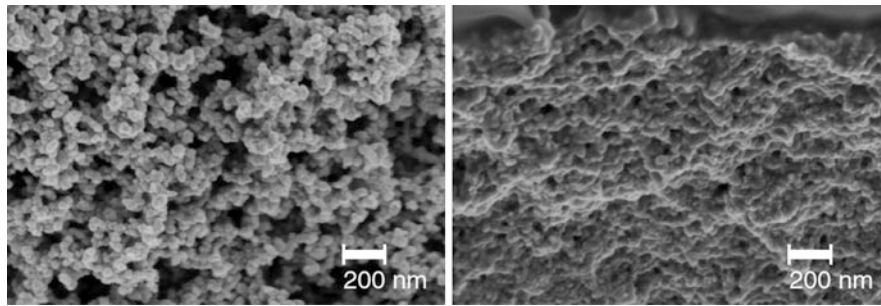
PEM Fuel Cells, Materials and Design Development Challenges. **Figure 7**

Scanning electron microscopy (SEM) cross section of an MEA of which 8% wt. of the cathode carbon-support had been corroded (see also Fig. 6). The initial cathode electrode thickness was identical to the anode electrode thickness shown in the picture (Reproduced from H.A. Gasteiger et al. [3] by permission from Springer)

of carbon support-corrosion was measured by on-line monitoring of the CO<sub>2</sub> formation rate).

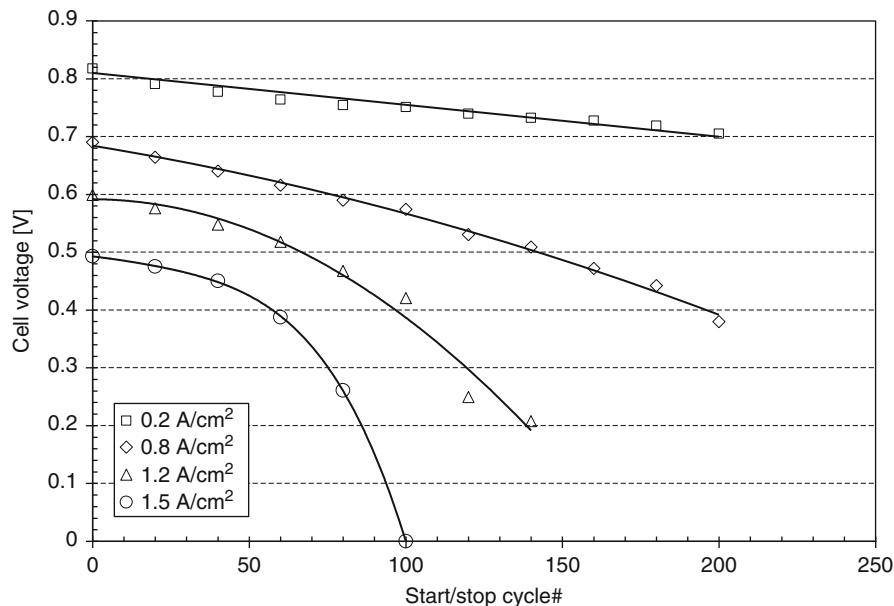
As shown in Fig. 7, the extent of cathode thinning can be monitored by cross section measurements via SEM or optical microscopy [23]. It is accompanied by a loss of electrode void volume as is illustrated in Fig. 8, showing high-resolution SEM cross sections of a nondegraded cathode electrode (left-hand-side) and of a degraded cathode (right-hand-side).

The fundamental start/stop mechanism was first reported by Reiser et al. [11], and occurs when one part of the anode flow-field is partially filled with hydrogen and another part is filled with air, a situation which occurs during the start-up of a fuel cell (hydrogen displacing air in the anode flow-field) or during shutdown (air diffusing into a hydrogen filled anode flow-field after the hydrogen supply is shut off). Detailed explanations of the processes which lead to very high voltages on the cathode electrode ( $\gg 1$  V vs.



PEM Fuel Cells, Materials and Design Development Challenges. Figure 8

SEM micrographs of freeze fractured sections of the cathode electrodes of MEAs. *Left*: nondegraded MEA; *right*: MEA aged by localized hydrogen starvation. SEM analysis was done without mounting of the MEAs in epoxy (Reproduced from R.N. Carter et al. [23] by permission from John Wiley & Sons)



PEM Fuel Cells, Materials and Design Development Challenges. Figure 9

Cell-voltage decays for different current densities as a function of start/stop cycles for an MEA with a platinum catalyst supported on a conventional carbon-support. Conditions: H<sub>2</sub>/air (66% inlet RH) at 150 kPa<sub>abs</sub> and 80°C, aged at a H<sub>2</sub>/air-front residence time of 1.3 s (Reproduced from P. T. Yu et al. [12] by permission from The Electrochemical Society)

RHE) have been given elsewhere [11, 34] and the interested reader is referred to these references. The voltage degradation rates produced by start/stop can be simulated conveniently by sending H<sub>2</sub>/air fronts through the anode flow-field of a fuel cell (single cell or stack), whereby the voltage degradation rates are proportional to the residence time of the H<sub>2</sub>/air front and increase with increasing current density [12], as would be

expected for gas-transport induced voltage losses. Such an experiment is shown in Fig. 9, whereby the H<sub>2</sub>/air-front residence time of 1.3 s is roughly 10 times longer than that which can be achieved during the start-up of a fuel cell stack (residence times much shorter than 0.1 s are typically not achievable due to engineering constraints). Under these conditions, the cell voltage at 1.5 A cm<sup>-2</sup> decreases to 0 V within only

100 cycles! Under normal start-up conditions, the residence time would be roughly 10 times shorter, so that a maximum of 1,000 start-up cycles could be performed, which is far short from the automotive target of  $\approx 40,000$  starts during the life time of a vehicle.

Therefore, mitigation strategies had to be devised and implemented. Currently, most mitigation strategies are based on system design (short residence times, stack storage under hydrogen, cell shorting, etc. [13]), but on the long-term, additional materials-based mitigation strategies are required. These include implementation of graphitized carbon-supports [12, 44, 45] or of more corrosion-resistant alternative support materials, lowering of the anode catalyst loading which reduces the ORR activity of the anode electrode or development of anode catalysts with low ORR activity [14], or the incorporation of highly active oxygen evolution catalysts in the cathode electrode. Again, the reader is referred to the literature for a detailed discussion [12, 44, 45].

A mechanism very closely related to the start/stop degradation is the so-called localized hydrogen starvation. It was first discussed by Patterson and Darling [9], and occurs when large sections of the anode flow-field are not being supplied with hydrogen due to blockage of the flow-field channel or of the diffusion medium by liquid water. In that case, oxygen permeating through the membrane from the anode to the cathode side creates an analogous situation to that produced by a  $H_2$ /air front, viz., the simultaneous presence of hydrogen and oxygen in the anode. Consequently, cathode thinning is observed also in the case of localized hydrogen starvation [10, 23], albeit at a slower rate. While the systems mitigation strategies are different from those used in the case of start/stop, the materials mitigation strategies are identical, with one additional materials mitigation approach: since the maximum carbon corrosion rate is limited by the oxygen permeation rate through the membrane, ionomeric membranes with reduced oxygen permeability (typical for most hydrocarbon-based ionomers) would lower the damage by localized  $H_2$  starvation [10].

### Membrane Development Needs and Approaches

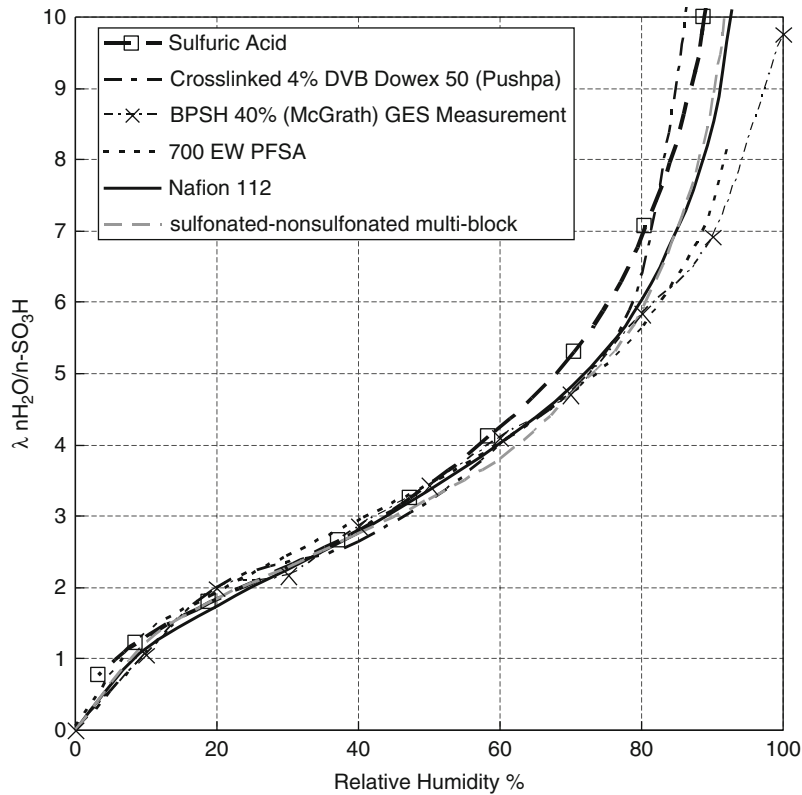
As indicated earlier, water is needed in the membrane to dissociate protons (i.e., making the material proton

conducting) from the acid functionality or protogenic groups and then secondly to establish a continuous pathway for long range proton transport. The water content of a PEM decreases with decreasing water activity. It has been observed [30] that the water uptake in sulfonic acid ( $-SO_3H$ )-based PEMs is essentially identical when plotted as function of the water content when the latter is expressed in terms of,  $\lambda$ , the number of water molecules per acid group (i.e.,  $\lambda = H_2O/SO_3H$ ) as seen in Fig. 10. It is also clear from this data that the isotherms of the various PEM are essentially identical to that of sulfuric acid at relatively low levels of hydration (i.e.,  $\lambda < 6$ ), suggesting that it is the enthalpy of hydration of the  $-SO_3H$  group which is driving water uptake in the low-RH region.

This experimental isotherm has been shown to fit either the Brunauer-Emmett-Teller model [46] or a simple empirical polynomial fit at a given temperature [47]. Finally, above  $\lambda \approx 6$  (i.e.,  $>80\%$  RH) the isotherms diverge, indicating that properties including polymer structure and morphology, charge density, cross-linking, etc., impact the absorption of water at high water content.

### Temperature Targets

For automotive application, it has been determined that a reasonable target for high temperature membrane operation is between  $110^\circ\text{C}$  and  $120^\circ\text{C}$  for  $H_2$  fueled fuel cell vehicles [5]. Heat rejection at this temperature with conventional packaging becomes feasible and the purity requirement for onboard  $H_2$  is reduced as the tolerance for CO improves to approximately 50 ppmv CO without air bleed at low anode catalyst loading ( $0.1\text{--}0.2 \text{ mg}_{\text{noble-metal}} \text{ cm}^{-2}$ ) [48]. Operation of stationary systems at  $140\text{--}160^\circ\text{C}$  with hydrocarbon-based  $H_2$  reformat would result in an increase in CO tolerance to about 0.1–0.5% allowing for a simpler or possibly no preferential oxidation (PROX) reactor. Although there is a small improvement in the oxygen reduction reaction kinetics if the system were operated at  $160^\circ\text{C}$ , this would be offset with a loss of about 70 mV in equilibrium voltage [49]. Furthermore, the strong specific adsorption of phosphate ions on platinum catalysts leads to a reduced ORR activity in the presence of phosphoric acid electrolyte if compared to sulfonic acid-based ionomers



PEM Fuel Cells, Materials and Design Development Challenges. Figure 10

Water uptake isotherms of various ionomers and sulfuric acid at 80°C. Dowex 50 is an ion-exchange resin made of 4% cross-linked polystyrene divinyl benzene; BPSH 40 is a 2-mil 40% randomly sulfonated biphenol sulfone; 700 EW PFSA is a 1-mil membrane with a structure similar to Nafion®; Nafion 112® is a 2-mil extruded membrane; and, PAEK triblock is a 1-mil triblock polyaryl ether ketone with a sulfonated middle block (Reproduced from C.K. Mittelsteadt and H. Liu. [30] by permission from John Wiley & Sons)

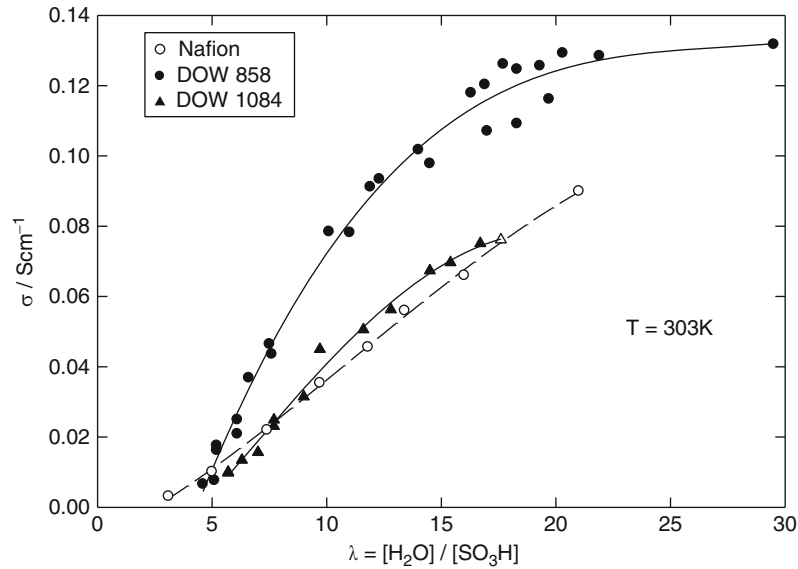
(see Table 4 and Fig. 20 in [50]). At temperatures >160°C, the stability of the carbon-support material is compromised [51].

### Proton Conductivity

Figure 11 displays the results of a seemingly rather subtle change in the length of the side chain in PFSA membranes: the Dow ionomers have a shorter perfluoroether side chain ( $-\text{OCF}_2\text{CF}_2\text{SO}_3\text{H}$ ) than Nafion ( $-\text{OCF}_2\text{CF}(\text{CF}_3)\text{OCF}_2\text{CF}_2\text{SO}_3\text{H}$ ). This plot clearly indicates that the density of the sulfonic acid groups as realized through alteration of the equivalent weight (EW, the grams of polymer per mole of acid) may bring about a fairly large change in the proton

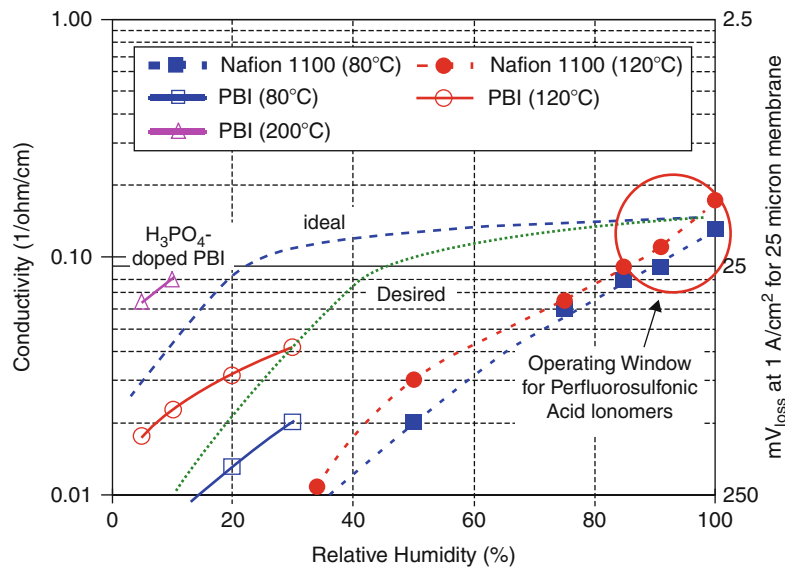
conductivity across the majority of the hydration isotherm. The proton conductivity of both ionomers, however, drops precipitously at water contents below 5  $\text{H}_2\text{O}/\text{SO}_3\text{H}$ .

The dependence of proton conductivity on RH for 1100 EW Nafion at 80°C and 120°C is shown in Fig. 12 and indicates that although the conductivity falls from about 0.10 to 0.01  $\text{S cm}^{-1}$  as the relative humidity is decreased from 100% to  $\approx 30\%$  it is essentially independent of temperature [5]. This is in stark contrast to polybenzimidazole (PBI) and phosphoric acid-doped PBI (also plotted in Fig. 12) which although exhibiting the typical fall in proton conductivity as the hydration level is decreased, show significantly higher conductivity at elevated temperatures [54]. Although the PBI



PEM Fuel Cells, Materials and Design Development Challenges. Figure 11

Room temperature proton conductivity of the short side chain (SSC) PFSA ionomer at 2 different (i.e., Dow 858 and Dow 1084  $g_{\text{ionomer}}$  per  $\text{mol}_{\text{H}^+}$ ) equivalent weights (EW) and Nafion as a function of water content expressed as  $\lambda = [\text{H}_2\text{O}] / [\text{SO}_3\text{H}]$ . The data clearly shows the significant effect the equivalent weight has on proton conductivity with the Dow 858 exhibiting conductivity twice that of the higher EW PFSA's (Reproduced from K. D. Kreuer et al. [52] by Elsevier Science)



PEM Fuel Cells, Materials and Design Development Challenges. Figure 12

Relationship between proton conductivity and adjoining gas stream humidity at various temperatures for Nafion (1100 EW) [5] and phosphoric acid-doped polybenzimidazole (PBI) [54]. The data clearly demonstrate that the increase in temperature from  $80^\circ\text{C}$  to  $120^\circ\text{C}$  has little effect on the conductivity of Nafion but a significant effect on the PBI systems. A curve is also shown for a material exhibiting the desired conductivity as a function of the relative humidity that would be ideal for system simplification (The figure is reproduced from Gasteiger and Mathias [5] with permission from The Electrochemical Society)

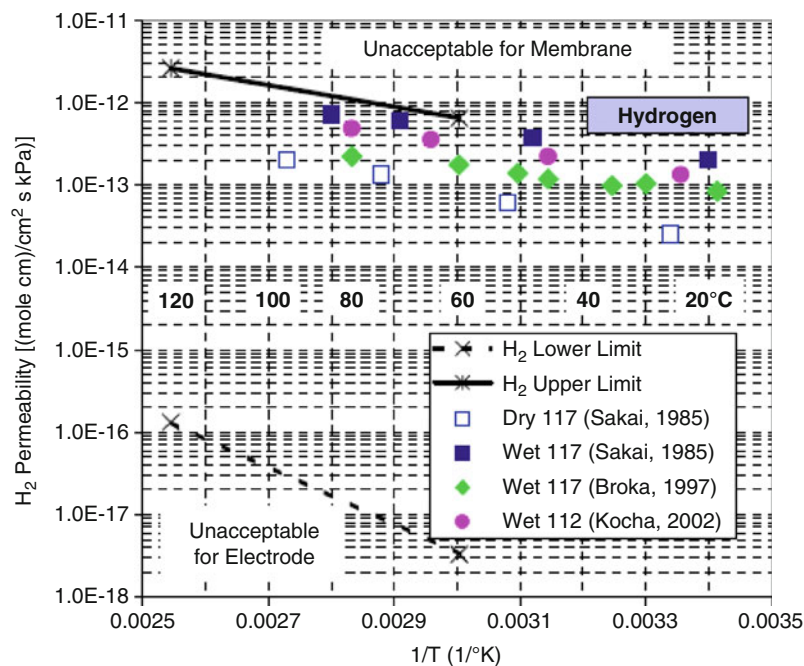
systems appear to offer the promise of high temperature fuel cell operation, they suffer from other issues including adsorption on the surfaces of platinum-based catalysts and the leaching of phosphoric acid from the electrolyte. Although a significant body of research has focused on determining the hydrated morphology of PFSA membranes, the connections between the structure and morphology with the transport properties are not fully understood. The microstructure of the PFSA polymer not only affects the proton conductivity but also other properties including methanol permeability (i.e., for direct methanol fuel cells), water diffusion, and electro-osmotic drag.

### Reactant Gas Permeability

Other important properties of the ionomer include the permeability to both  $H_2$  and  $O_2$  gas. The PEM must not be too permeable to the reactive gases, as excessive gas crossover through the membrane would result in fuel efficiency losses. However, the ionomer in the electrodes

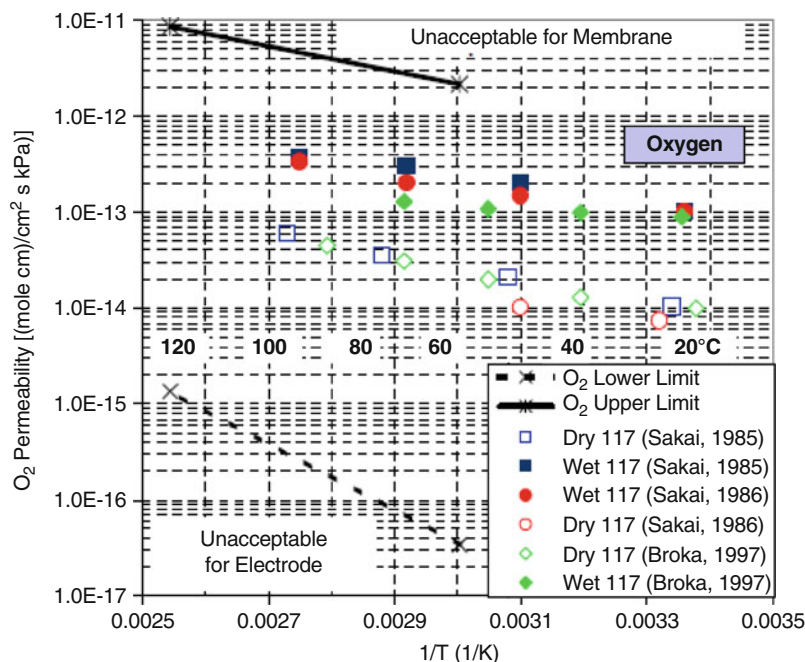
must possess sufficient permeability to allow transport of the reactant without imposing any significant concentration gradients and/or mass transfer losses.

The rate of the permeation of gases in PFSA membranes is proportional to the product of the permeability coefficient and the partial pressure (i.e., the driving force) and inversely proportional to the membrane thickness. The permeation leads to a fuel cell crossover efficiency loss due to two components: hydrogen consumption at the anode from  $O_2$  crossover and hydrogen loss to the cathode and subsequent reaction with oxygen. A relatively straightforward analysis [5] based on fuel cell operation with instantaneously varying loads of  $0.05 \text{ A cm}^{-2}$  at  $60^\circ\text{C}$  and  $2 \text{ A cm}^{-2}$  at  $120^\circ\text{C}$  has estimated both upper and lower limits for the permeability of  $H_2$  and  $O_2$  which are shown as a function of temperature in Figs. 13 and 14, respectively. The permeability window for both reactant gases is approximately 3–4 orders of magnitude. Hence, the permeability coefficient,  $k$ , for  $H_2$  transport must fall within the range:  $1 \times 10^{-17} < k_{H_2} < 1 \times 10^{-12} \text{ mol cm}$



PEM Fuel Cells, Materials and Design Development Challenges. Figure 13

Hydrogen gas permeability as a function of both temperature and relative humidity. Upper limit (solid line) defined by crossover losses (assuming no contribution from  $O_2$  crossover), lower limit (dotted line) defined by the transport requirements of the ionomer in the electrode. Data for both dry and wet Nafion 1100 EW membranes taken from Refs. [55–57] (Reproduced from H.A. Gasteiger and M.F. Mathias [5] with permission from The Electrochemical Society)



PEM Fuel Cells, Materials and Design Development Challenges. Figure 14

Oxygen gas permeability as a function of both temperature and relative humidity. Upper limit (*solid line*) defined by crossover losses (assuming no contribution from  $H_2$  crossover), lower limit (*dotted line*) defined by the transport requirements of the ionomer in the electrode. Data for both dry and wet Nafion 1100 EW membranes taken from Refs. [55–57] (Reproduced from H.A. Gasteiger and M.F. Mathias [5] with permission from The Electrochemical Society)

$cm^{-2} s^{-1} kPa^{-1}$ , and for  $O_2$  transport within the range:  $1 \times 10^{-16} < k_{H_2} < 3 \times 10^{-12} mol cm cm^{-2} s^{-1} kPa^{-1}$ , where the lower limits are dictated by the ionomer in the electrodes and the upper limits by the ionomer in the membrane.

### Morphology

As indicated earlier, all presently available PEMs must be humidified in order to exhibit sufficient proton conductivity to function effectively as the electrolyte in a fuel cell. When humidified, the PEMs swell with the absorbed water resulting in a hydrated morphology where the aqueous phase is confined within the polymer to domains that are typically only a few nanometers in dimensions [56–62]. This morphology and the interactions driving their formation are key to the understanding of the morphological stability of the ionomer and transport properties [63, 64]. Both are of paramount importance for the application of

such materials in PEM fuel cells: morphological stability under operating conditions is not only important for keeping the integrity of the ionomer membrane but also for a stable microstructure of the active electrode layers usually containing a high volume fraction of ionomer. Proton transport is actually a very complex phenomenon [36, 63, 65] comprising different species (protonic charge carriers, water, and gases dissolved therein) and different transport modes (diffusional, hydrodynamic), but the key process here remains the proton conductivity.

The microstructure is usually the consequence of a constrained hydrophobic/hydrophilic separation, and this can be controlled by adjusting the concentration of protogenic groups (ion-exchange capacity IEC). A high IEC corresponds to a high charge carrier concentration and generally leads to a high uptake of water, both being favorable for high proton conductivity. However, the high IEC leads to severe swelling which goes along with a loss of morphological instability and

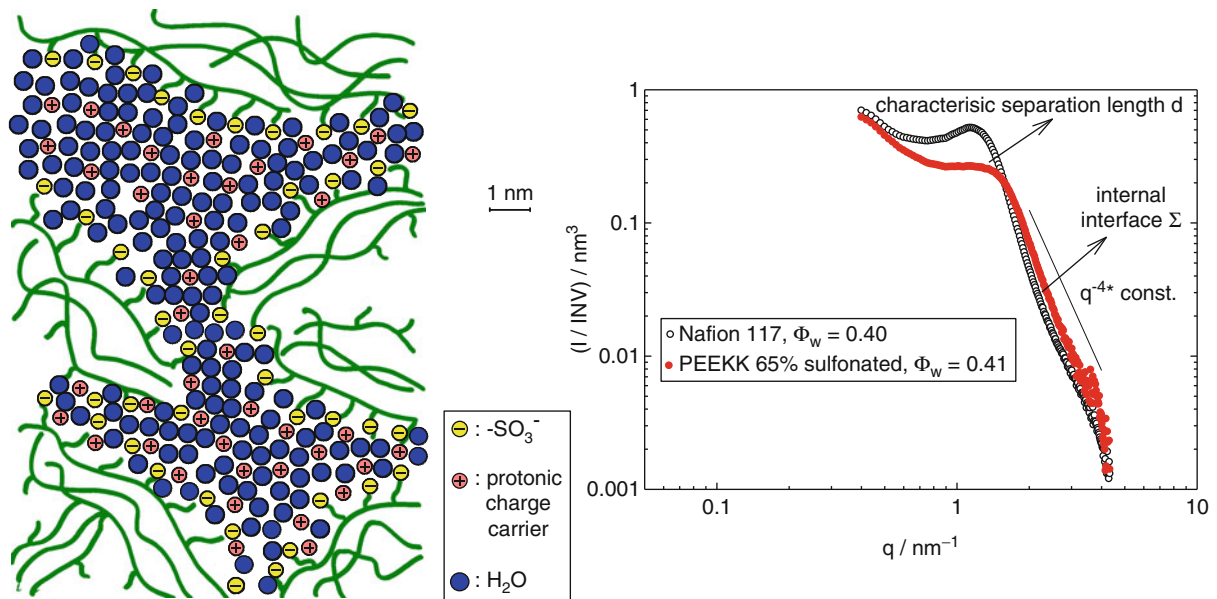


eventually leads to a complete dissolution of the ionomer. Therefore, the optimum IEC is essentially a tradeoff between these two properties.

Interestingly, this is different for different types of ionomers which is because of the differences in the backbone/backbone interactions and hydrated morphology. Such morphological details are usually obtained from small angle X-ray (SAXS) diffraction experiments (see Fig. 15). A recent interpretation of the SAXS spectra of fully swollen Nafion has attracted considerable attention [62]. The authors claim the existence of parallel cylindrical micelles bearing the water and the dissociated protons. Such a morphology could easily explain the relatively high conductivity of PFSA membranes at low water contents corresponding to a low IEC and low degree of hydration (low RH): cylindrical structures provide high connectivity within the water structure even for low water volume fractions and a high local proton mobility because of their relatively large width compared to water structures of higher dimensionality (a cylinder

width of 2.4 nm is suggested which is large compared to about 1 nm suggested by other models). But it should be noted that other groups have raised serious doubts about this model [67, 68]. It has recently been proposed that the morphology resembles two dimensional water structures, and that it is essentially the tortuosity of locally flat water structures which determines percolation and therefore proton conductivity in such materials.

In any case, there is a clear difference in the morphology of PFSA and hydrocarbon membranes which was first pointed out in a comparative SAXS study [60]. While proton conductivity at high levels of hydration is mainly dependent on the IEC only, the decrease in proton conductivity with decreasing water content is more severe in hydrocarbon membranes. The reason for this characteristic difference is suggested to be the more pronounced hydrophobic/hydrophilic separation of PFSA ionomers which leads to a better connectivity within the aqueous domain and locally to more bulk-like properties of the water of hydration. But there are



PEM Fuel Cells, Materials and Design Development Challenges. **Figure 15**

*Left:* A schematic representation of the fully hydrated morphology of a PFSA ionomer (e.g., Nafion) under the assumptions of a cubic lattice model which fitted data from small angle X-ray scattering (SAXS) experiments. *Right:* SAXS spectra of hydrated Nafion and a hydrated sulfonated polyetherketone. The characteristic hydrophobic/hydrophilic separation lengths are obtained from the position of the ionomer peaks while the internal hydrophobic/hydrophilic interfaces are obtained from the intensities in the Porod regime. First reported in Ref. [66]

also characteristic disadvantages of PFSA membranes: (1) because of the bulk-like properties of the hydration water, the transport of water has a large hydrodynamic component which shows up as large electro-osmotic water drag and high water/gas permeation coefficients and (2) the viscoelastic properties are severely decaying with temperature. At this stage, it should be noted that so-called short side chain ionomers show a slightly weaker hydrophobic/hydrophilic separation which actually reduces the conductivity for a given IEC. But the significantly higher morphological stability especially at higher temperature (higher  $T_g$ ) allows for significantly higher IECs without significantly compromising the elastic properties (storage modulus). PFSA membranes with shorter side chains (e.g., Dow, 3 M, Aquivion) seem to be a real improvement over the traditional long side chain PFSA membrane Nafion [52].

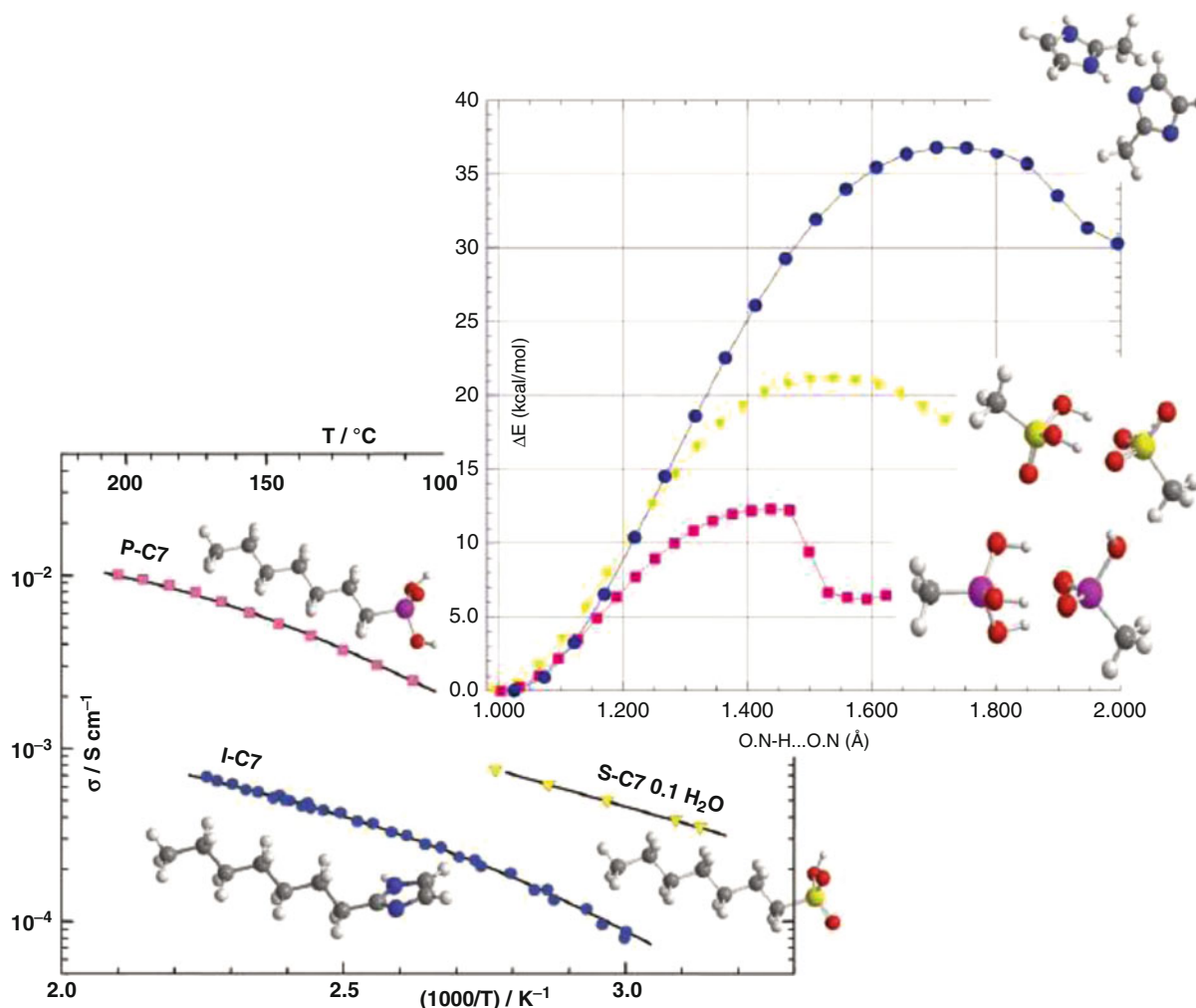
Despite the disadvantageous conductivity behavior at low levels of hydration, the other alternatives are hydrocarbon membranes. They usually do not show softening in the temperature range of fuel cell operation and the higher dispersion of the hydration water (smaller width of the water domain) leads to significantly lower hydrodynamic water transport. The lower conductivity can actually be increased by introducing another heterogeneity with respect to the degree of sulfonation (IEC) on the 10–100 nm scale. This can lead to a significant increase of proton conductivity provided that bi-continuous morphologies are formed. The reason is the highly nonlinear increase of proton conductivity with increasing IEC. An interesting approach toward such morphologies is the formation of multiblock copolymers consisting of an alternating sequence of highly sulfonated hydrocarbon segments and unsulfonated segments [69]. When cast from solutions, such polymers undergo a constrained microphase separation. The hydrophilic sulfonated phase then still provides a sufficiently high conductivity while the nonsulfonated phase can give the material morphological stability and reduces swelling to an acceptable level (see also the entry in this encyclopedia: “► [Membrane Electrolytes: from Perfluoro Sulfonic Acid to Hydrocarbon Ionomers](#)” by Miyatake).

Optimizing and controlling the microstructure of such materials is one of the major materials design issues. Here, the main challenges are to maximize the

local concentration of sulfonic acid functions within the hydrophilic phase and to obtain morphological stability with relatively small volume fractions of unsulfonated phase. A recent study on the model system of aqueous methylsulfonic acid ( $\text{CH}_3\text{SO}_3\text{H}(\text{aq})$ ) clearly demonstrates that there is still a huge potential for increasing proton conductivity at low relative humidity [70].

### Choice of the Protogenic Group

The source of the protons in current state-of-the-art PEM fuel cells is the highly acidic sulfonic acid functional group. As pointed out above, hydration is required to dissociate the protons of these groups and to mobilize the protonic charge carriers through solvation (hydration). Since the hydration requirement is one of the issues limiting the operation temperature of sulfonic acid-based electrolytes, there has been an extensive search for other protogenic groups which may enable high proton conductivity at lower water activities [71, 72]. One of the most interesting approaches is to use functional groups which are amphoteric in the sense that they may act both as a proton donor and as a proton acceptor. Conceptually, they can combine the role of the proton source and the proton solvent in one functional group: high self-dissociation may lead to a sufficiently high charge carrier concentration and structural diffusion (i.e., proton hopping or shuttling) within a dynamically disordered hydrogen bond network and thereby provide high charge carrier mobility. The concept has been proven for imidazole [73] and phosphonic functionalized model compounds [74]. A comparative study, however, of sulfonic acid, phosphonic acid, and imidazole as protogenic groups suggests that only phosphonic acid has some potential to substitute for the sulfonic acid functional group [75, 76]. Experimentally determined proton conductivities of functionalized heptanes examined in this study [75] and corresponding computed proton transfer barriers are shown in Fig. 16. This study actually investigated not only transport but also stability issues and the participation of these groups in the electrochemical reactions. Apart from the fact that the imidazole functionalized oligomer showed the lowest proton conductivity, reaction with oxygen was observed and the platinum catalyst was



PEM Fuel Cells, Materials and Design Development Challenges. Figure 16

A compound figure consisting of both experimental [75] and theoretical results [76]. Lower Left: Measured proton conductivities under dry conditions versus temperature of three monofunctionalized heptanes: 1-heptylphosphonic acid (P-C7), magenta squares; 1-heptylsulfonic acid (S-C7), yellow triangles; and 2-heptylimidazole (I-C7), blue circles. Upper Right: Computed energetic barriers for neat (i.e., acid to acid) proton transfer for methylphosphonic acid, magenta squares; methylsulfonic acid, yellow triangles; and methylimidazole, blue circles. The combined results suggest that proton conductivity is at least partially a function of the barrier for proton transfer: the experimental proton conductivities are inversely related to the computed proton transfer barrier

blocked at high potentials where oxygen reduction takes place. The phosphonic acid-based system showed a clear signature of proton conductivity in the “water free” state, but some water activity was essential to prevent condensation reactions which immediately suppress proton conductivity. It appears that the

condensation issue is more severe for systems with immobilized phosphonic acid groups compared to pure liquid phosphonic and phosphoric acid, and it remains a challenge to immobilize phosphonic acid functional groups without increasing the susceptibility for condensation [77].

## Future Directions

Despite the significant and substantial progress in PEM fuel cell technology achieved during the past couple of decades, the large-scale market introduction of this technology into applications such as vehicular power will require overcoming the high costs associated with the components of the fuel cell stack (i.e., the anode and cathode catalysts, the ionomeric membrane, bipolar plates, etc.). The platinum or platinum-alloy based electrodes in current state-of-the-art fuel cells will constitute a significant fraction of the overall cost of a PEMFC stack if produced at large number. The precious metal catalysts are the only component in the fuel cell stack that will not benefit from an economies of scale, and hence the research and development of nonprecious metal catalysts is an important challenge that must be overcome.

Although it would be best if the platinum catalysts were replaced at both electrodes, the reduction of oxygen (at the cathode) requires much more Pt and hence the development of nonnoble metal catalysts with sufficient and durable ORR activity is a major focus of current and ongoing research [78–80]. Significant effort and progress has recently been achieved toward the development of catalysts with nitrogen coordinated with either iron or cobalt in a carbon matrix or support (i.e., Fe/N/C or Co/N/C). These systems are showing great promise with turnover frequencies comparable to that of current Pt/C catalysts. A major challenge (and still largely unexplored) is the stability and durability of these catalysts in the hostile electrochemical environment of an operating fuel cell [81].

Although substantial progress has been made on the development of advanced high performance PEMs, there are currently no ionomers that exhibit sufficiently high proton conductivity and durability under hot (i.e.,  $>100^{\circ}\text{C}$ ) and “dry” ( $<30\%$  RH) conditions. However, there are several potentially promising routes that may ultimately lead to electrolytes exhibiting both high chemical and thermal stability and not requiring humidification. Increasing the density of sulfonic acid groups (i.e., by lowering the IEC) in PFSA ionomers through shortening the side chain and/or having tethering more than one protogenic group per side chain warrants further research. The development of the highly sulfonated polysulfones also seems

to offer some promise as these materials demonstrate that with interpenetrating aqueous domains of very small diameters the proton conductivity is still very high. This class of materials, however, must be made with lower water solubility in water and with resistance to breakage through elongation.

## Bibliography

1. Gasteiger HA, Marković NM (2009) Just a dream – or future reality? Advances in catalyst development offer hope for commercially viable hydrogen fuel cells. *Science* 324:48–49
2. Mathias MF, Makharia R, Gasteiger HA, Conley JJ, Fuller T, Gittleman C, Kocha SS, Miller D, Mittelsteadt C, Xie T, Yan SG, Yu PT (2005) Two fuel cell cars in every garage? *Electrochem Soc Interface* 14(2):24–35, Pennington
3. Gasteiger HA, Gu W, Litteer B, Makharia R, Brady B, Budinski M, Thompson E, Wagner FT, Yan SG, Yu PT (2007) Catalyst degradation mechanisms in PEM and direct methanol fuel cells. In: Kakac S, Pramanjaroenkij A, Vasiliev L (eds) Proceedings of the conference of the NATO-Advanced-study-institute on mini-micro fuel cells – fundamentals and applications, Cesme Izmir, 22 July–03 Aug. Springer, Dordrecht, pp 225–233
4. Masten DA, Bosco AD (2003) System design for vehicle applications – GM/Opel. In: Vielstich W, Lamm A, Gasteiger HA (eds) Handbook of fuel cells – fundamentals, technology and applications, vol 4. Wiley, Chichester, pp 714–724
5. Gasteiger HA, Mathias MF (2005) Fundamental research and development challenges in polymer electrolyte fuel cell technology. In: Murthy M, Fuller TF, Van Zee JW (eds) Proceedings of the symposium on proton conducting membrane fuel cells III, 202nd ECS Meeting, held in Salt Lake City, Utah in the year 2002, vol PV 2002–31. The Electrochemical Society, Pennington, pp 1–24
6. Endoh E (2009) Highly durable PFSA membranes. In: Vielstich W, Gasteiger HA, Yokokawa H (eds) Handbook of fuel cells – advances in electrocatalysis, materials, diagnostics, and durability, vol 5. Wiley, Chichester, pp 361–374
7. Liu H, Coms FD, Zhang J, Gasteiger HA, LaConti AB (2009) Chemical degradation: correlations between electrolyzer and fuel cell findings. In: Büchi FN, Inaba M, Schmidt TJ (eds) Polymer electrolyte fuel cell durability. Springer, New York, pp 71–118
8. Lai YH, Dillard DA (2009) Mechanical durability characterization and modeling of ionomeric membranes. In: Vielstich W, Gasteiger HA, Yokokawa H (eds) Handbook of fuel cells – advances in electrocatalysis, materials, diagnostics, and durability, vol 5. Wiley, Chichester, pp 403–419
9. Patterson TW, Darling RM (2006) Damage to the cathode catalyst of a PEM fuel cell caused by localized fuel starvation. *Electrochem. Solid-State Lett.* 9:A183–A185
10. Gu W, Yu PT, Carter RN, Makharia R, Gasteiger HA (2010) Modeling of membrane-electrode assembly degradation in proton-exchange-membrane fuel cells – local  $\text{H}_2$

- starvation and start-stop induced carbon-support corrosion. In: Pasaogullari U, Wang C-Y (eds) Modeling and diagnostics of polymer electrolyte fuel cells, vol 49, Modern aspects of electrochemistry. Springer, New York, pp 45–85
11. Reiser CA, Bregoli L, Patterson TW, Yi JS, Yang JDL, Perry ML, Jarvi TD (2005) A reverse current decay mechanism for fuel cells. *Electrochem. Solid-State Lett.* 8:A273–A276
  12. Yu P, Gu W, Makharia R, Wagner FT, Gasteiger HA (2006) The impact of carbon stability on PEM fuel cell startup and shut-down voltage degradation. *ECS Trans* 3:797–809
  13. Perry ML, Patterson TW, Reiser C (2006) System strategies to mitigate carbon corrosion in fuel cells. *ECS Trans* 3: 783–795
  14. Genorio B, Subbaraman R, Strmcnik D, Tripkovic D, Stamenkovic VR, Marković NM (2011) Tailoring the selectivity and stability of chemically modified platinum nanocatalysts to design highly durable anodes for PEM fuel cells. *Angew Chem Int Ed* 50:1–6
  15. Ralph TR, Hudson S, Wilkinson DP (2006) Electrocatalyst stability in PEMFCs and the role of fuel starvation and cell reversal tolerant anodes. *ECS Trans* 1:67–84
  16. Debe MK, Schmoeckel AK, Vernstrom GD, Atanasoski R (2006) High voltage stability of nanostructured thin film catalysts for PEM fuel cells. *J Power Sources* 161:1002–1011
  17. Carter RN, Greszler TA, Baker DR (2009) Technique for measuring gas transport resistance in application-scale aged fuel cell gas diffusion media. *ECS Trans* 25:225–231
  18. Perry ML, Patterson T, Madden T (2010) GDL degradation in PEFC. *ECS Trans* 33:1081–1087
  19. Stephens IEL, Bondarenko AS, Perez-Alonso FJ, Calle-Vallejo F, Bech L, Johansson TP, Jepsen AK, Frydenal R, Knudsen BP, Rossmeisl J, Chorkendorff I (2011) Tuning the activity of Pt (111) for oxygen electroreduction by subsurface alloying. *J Am Chem Soc* 133:5485–5491
  20. Gu W, Baker DR, Liu Y, Gasteiger HA (2009) Proton exchange membrane fuel cell (PEMFC) down-the-channel performance model. In: Vielstich W, Gasteiger HA, Yokokawa H (eds) Handbook of fuel cells – advances in electrocatalysis, materials, diagnostics, and durability, vol 5. Wiley, Chichester, pp 631–657
  21. Pasaogullari U (2009) Heat and water transport models for polymer electrolyte fuel cells. In: Vielstich W, Gasteiger HA, Yokokawa H (eds) Handbook of fuel cells – advances in electrocatalysis, materials, diagnostics, and durability, vol 5. Wiley, Chichester, pp 616–630
  22. Freunberger SA, Reum M, Büchi FN (2009) Design approaches for determining local current and membrane resistance in polymer electrolyte fuel cells (PEFCs). In: Vielstich W, Gasteiger HA, Yokokawa H (eds) Handbook of fuel cells – advances in electrocatalysis, materials, diagnostics, and durability, vol 5. Wiley, Chichester, pp 603–615
  23. Carter RN, Gu W, Brady B, Yu PT, Subramanian K, Gasteiger HA (2009) Membrane electrode assembly (MEA) degradation mechanism studies by current distribution measurements. In: Vielstich W, Gasteiger HA, Yokokawa H (eds) Handbook of fuel cells – advances in electrocatalysis, materials, diagnostics, and durability, vol 5. Wiley, Chichester, pp 829–843
  24. Trabold TA, Owejan JP, Gagliardo JJ, Jacobson DL, Hussey DS, Arif M (2009) Use of neutron imaging for proton exchange membrane fuel cell (PEMFC) performance analysis and design. In: Vielstich W, Gasteiger HA, Yokokawa H (eds) Handbook of fuel cells – advances in electrocatalysis, materials, diagnostics, and durability, vol 5. Wiley, Chichester, pp 658–672
  25. Wipke K, Sprick S, Kurtz J, Garbak J (2009) Field experience with fuel cell vehicles. In: Vielstich W, Gasteiger HA, Yokokawa H (eds) Handbook of fuel cells – advances in electrocatalysis, materials, diagnostics, and durability, vol 6. Wiley, Chichester, pp 893–904
  26. Thompson EL, Gu W, Gasteiger HA (2009) Performance during start-up of proton exchange membrane (PEM) fuel cells at subfreezing conditions. In: Vielstich W, Gasteiger HA, Yokokawa H (eds) Handbook of fuel cells – advances in electrocatalysis, materials, diagnostics, and durability, vol 5. Wiley, Chichester, pp 699–717
  27. Mathias M, Roth J, Fleming J, Lehnert W (2003) Diffusion media materials and characterization. In: Vielstich W, Lamm A, Gasteiger HA (eds) Handbook of fuel cells – fundamentals, technology and applications, vol 3. Wiley, Chichester, pp 517–537
  28. Baker DR, Caulk DA, Neyerlin KC, Murphy MW (2009) Measurement of oxygen transport resistance in PEM fuel cells by limiting current methods. *J Electrochem Soc* 156: B991–B1003
  29. Caulk DA, Baker DR (2011) Modeling two-phase water transport in hydrophobic diffusion media for PEM fuel cells. *J Electrochem Soc* 158:B384–B393
  30. Mittelsteadt CK, Liu H (2009) Conductivity, permeability, and ohmic shorting of ionomeric membranes. In: Vielstich W, Gasteiger HA, Yokokawa H (eds) Handbook of fuel cells – advances in electrocatalysis, materials, diagnostics, and durability, vol 5. Wiley, Chichester, pp 345–359
  31. Liu Y, Ji C, Gu W, Jorne J, Gasteiger HA (2011) Effects of catalyst carbon support on proton conduction and cathode performance in PEM fuel cells. *J Electrochem Soc* 158: B614–B621
  32. Neyerlin KC, Gu W, Jorne J, Clark A, Gasteiger HA (2007) Cathode catalyst utilization for the ORR in a PEMFC – analytical model and experimental validation. *J Electrochem Soc* 154: B279–B287
  33. Neyerlin KC, Gu W, Jorne J, Gasteiger HA (2007) Study of the exchange current density for the hydrogen oxidation and evolution reactions. *J Electrochem Soc* 154:B631–B635
  34. Gasteiger HA, Baker DR, Carter RN, Gu W, Liu Y, Wagner FT, Yu PT (2010) Electrocatalysis and catalyst degradation challenges in proton exchange membrane fuel cells. In: Stolten D (ed) Hydrogen and fuel cells. fundamentals, technologies, and applications. Wiley-VCH, Weinheim, pp 3–16
  35. Jaouen F, Proietti E, Lefèvre M, Chenitz R, Dodelet J-P, Wu G, Chung HT, Johnston CM, Zelenay P (2011) Recent advances in

- non-precious metal catalysis for oxygen reduction reaction in polymer electrolyte fuel cells. *Energy Environ Sci* 4:114–130
36. Paddison SJ (2003) Proton conduction mechanisms at low degrees of hydration in sulfonic acid-based polymer electrolyte membranes. *Annu Rev Mater Res* 33:289–319
  37. Ferreira PJ, la O' GJ, Shao-Horn Y, Morgan D, Makharia R, Kocha SS, Gasteiger HA (2005) Instability of Pt/C electrocatalysts in proton exchange membrane fuel cells: a mechanistic investigation. *J Electrochem Soc* 152:A2256–A2271
  38. Kinoshita K, Lundquist JT, Stonehart P (1973) Potential cycling effects on platinum electrocatalyst surfaces. *J Electroanal Chem* 48:157–166
  39. Kawahara S, Mitsuhashi S, Ota K-I, Kamiya N (2006) Deterioration of Pt catalyst under potential cycling. *ECS Trans* 3(1): 625–631
  40. Zhang J, Litteer BA, Gu W, Liu H, Gasteiger HA (2007) Effect of hydrogen and oxygen partial pressure on Pt precipitation within the membrane of PEMFCs. *J Electrochem Soc* 154: B1006–B1011
  41. Wagner FT, Yan SG, Yu PT (2009) Catalyst and catalyst-support durability. In: Vielstich W, Gasteiger HA, Yokokawa H (eds) *Handbook of fuel cells – advances in electrocatalysis, materials, diagnostics, and durability*, vol 5. Wiley, Chichester, pp 250–263
  42. Shao-Horn Y, Sheng WC, Chen S, Ferreira PJ, Holby EF, Morgan D (2007) Instability of supported platinum nanoparticles in low-temperature fuel cells. *Top Catal* 46:285–305
  43. Chen S, Gasteiger HA, Hayakawa K, Tada T, Shao-Horn Y (2010) Platinum-Alloy catalyst degradation in proton exchange membrane fuel cells: nanometer-scale compositional and morphological changes. *J Electrochem Soc* 157:A82–A97
  44. Yu PT, Gu W, Zhang J, Makharia R, Wagner FT, Gasteiger HA (2009) Carbon-support requirements for highly durable fuel cell operation. In: Büchi FN, Inaba M, Schmidt TJ (eds) *Polymer electrolyte fuel cell durability*. Springer, New York, pp 29–53
  45. Gallagher KG, Darling RM, Fuller TF (2009) Carbon-support corrosion mechanisms and models. In: Vielstich W, Gasteiger HA, Yokokawa H (eds) *Handbook of fuel cells – advances in electrocatalysis, materials, diagnostics, and durability*, vol 5. Wiley, Chichester, pp 819–828
  46. Thampan T, Malhorta S, Tang H, Datta R (2000) Modeling of conductive transport in proton-exchange membranes for fuel cells. *J Electrochem Soc* 147:3242–3250
  47. Springer TE, Zawodzinski T, Gottesfeld S (1991) Polymer electrolyte fuel cell model. *J Electrochem Soc* 138:2334–2342
  48. Springer T, Zawodzinski T, Gottesfeld S (1997) In: McBreen J, Mukerjee S, Srinivasan S (eds) *Electrode materials and processes for energy conversion and storage*, Proceedings series, vol PV 97–13. The Electrochemical Society, Pennington, pp 15–24
  49. Bernardi DM, Verbrugge MW (1992) A mathematical model of the solid-polymer-electrolyte fuel-cell. *J Electrochem Soc* 139:2477–2491
  50. Gasteiger HA, Garche J (2008) Fuel cells. In: Ertl G, Knözinger H, Schüth F, Weitkamp J (eds) *Handbook of heterogeneous catalysis*, 2nd edn. Wiley-VCH, Weinheim, pp 3081–3121
  51. Breault RD (2003) PAFC stack materials and stack design. In: Vielstich W, Lamm A, Gasteiger HA (eds) *Handbook of fuel cells: fundamentals, technology, and applications*, vol 3. Wiley, Chichester
  52. Kreuer KD, Schuster M, Obliers B, Diat O, Traub U, Fuchs A, Klock U, Paddison SJ, Maier J (2008) Short-side-chain proton conducting perfluorosulfonic acid ionomers: why they perform better in PEM fuel cells. *J Power Sources* 178: 499–509
  53. Alberti G, Casciola M, Massinelli L, Bauer B (2001) Polymeric proton conducting membranes for medium temperature fuel cells (110–160°C). *J Membrane Sci* 185:73–81
  54. Wainwright JS, Litt MH, Savinell RF (2003) High-temperature membranes. In: Vielstich W, Lamm A, Gasteiger HA (eds) *Handbook of fuel cells – fundamentals, technology and applications*, vol 3. Wiley, Chichester, pp 436–446
  55. Sakai T, Takenaka H, Wakabayashi N, Kawami Y, Torikai E (1985) Gas permeation properties of solid polymer electrolyte (SPE) membranes. *J Electrochem Soc* 132:1328–1332
  56. Broka K, Ekdunge P (1997) Oxygen and hydrogen permeation properties and water uptake of Nafion(R) 117 membrane and recast film for PEM fuel cell. *J Appl Electrochem* 27:281–289
  57. Kocha SS (2003) Principles of MEA preparation. In: Vielstich W, Lamm A, Gasteiger HA (eds) *Handbook of fuel cells – fundamentals, technology and applications*, vol 3. Wiley, Chichester, p 538
  58. Hsu WY, Gierke TD (1982) Elastic theory for ionic clustering in perfluorinated ionomers. *Macromolecules* 15:101–105
  59. Gebel G (2000) Structural evolution of water swollen perfluoro-sulfonated ionomers from dry membrane to solution. *Polymer* 41:5829–5838
  60. Kreuer KD (2001) On the development of proton conducting polymer membranes for hydrogen and methanol fuel cells. *J Membrane Sci* 185:29–39
  61. Rubatat L, Gebel G, Diat O (2004) Fibrillar structure of Nafion: matching fourier and real space studies of corresponding films and solutions. *Macromolecules* 37:7772–7783
  62. Schmidt-Rohr K, Chen Q (2008) Parallel cylindrical water nanochannels in Nafion fuel-cell membranes. *Nat Mater* 7:75–83
  63. Kreuer KD, Paddison SJ, Spohr E, Schuster M (2004) Transport in proton conductors for fuel-cell applications: simulations, elementary reactions, and phenomenology. *Chem Rev* 104:4637–4678
  64. Elliott JA, Paddison SJ (2007) Modelling of morphology and proton transport in PFSA membranes. *Phys Chem Chem Phys* 9:2602–2618
  65. Kreuer KD (2000) On the complexity of proton conduction phenomena. *Solid State Ionics* 136:149–160
  66. Ise M (2000) *Polymer Elektrolyt Membranen: Untersuchungen zur Mikrostruktur und zu den Transporteigenschaften für Protonen und Wasser*, PhD thesis, University of Stuttgart
  67. Kreuer KD (2011) *Advances in materials for proton exchange membrane fuel cells systems 2011*. Asilomar, Pacific Grove, 20–23 Feb 2011

68. Gebel G (2011) Advances in materials for proton exchange membrane fuel cells systems 2011. Asilomar, Pacific Grove, 20–23 Feb 2011
69. Hickner MA, Ghassemi H, Kim YS, Einsla BR, McGrath JE (2004) Alternative polymer systems for proton exchange membranes (PEMs). *Chem Rev* 104:4587–4612
70. Telfah A, Majer G, Kreuer KD, Schuster M, Maier J (2010) Formation and mobility of protonic charge carriers in methyl sulfonic acid–water mixtures: a model for sulfonic acid based ionomers at low degree of hydration. *Solid State Ionics* 181:461–465
71. Desmarteau DD (1995) Novel perfluorinated ionomers and ionenes. *J Fluorine Chem* 72:203–208
72. Schaberg MS, Abulu JE, Haugen GM, Emery MA, O’Conner SJ, Xiong PN, Hamrock SJ (2010) New multi acid side-chain ionomers for proton exchange membrane fuel cells. *ECS Trans* 33(1):609–627
73. Herz HG, Kreuer KD, Maier J, Scharfenberger G, Schuster MFH, Meyer WH (2003) New fully polymeric proton solvents with high proton mobility. *Electrochim Acta* 48: 2165–2171
74. Steininger H, Schuster M, Kreuer KD, Maier J (2006) Intermediate temperature proton conductors based on phosphonic acid functionalized oligosiloxanes. *Solid State Ionics* 177:2457–2462
75. Schuster M, Rager T, Noda A, Kreuer KD, Maier J (2005) About the choice of the protogenic group in PEM separator materials for intermediate temperature, low humidity operation: a critical comparison of sulfonic acid, phosphonic acid and imidazole functionalized model compounds. *Fuel Cells* 5: 355–365
76. Paddison SJ, Kreuer KD, Maier J (2006) About the choice of the protogenic group in polymer electrolyte membranes: ab initio modelling of sulfonic acid, phosphonic acid, and imidazole functionalized alkanes. *Phys Chem Chem Phys* 8: 4530–4542
77. Steininger H, Schuster M, Kreuer KD, Kaltbeitzel A, Bingöl B, Meyer WH, Schauff S, Brunklaus G, Maier J, Spiess HW (2007) Intermediate temperature proton conductors for PEM fuel cells based on phosphonic acid as protogenic group: a progress report. *Phys Chem Chem Phys* 9: 1764–1773
78. Bashyam R, Zelenay P (2006) A class of non-precious metal composite catalysts for fuel cells. *Nature* 443:63–66
79. Lefèvre M, Proietti E, Jaouen F, Dodelet JP (2009) Iron-based catalysts with improved oxygen reduction activity in polymer electrolyte fuel cells. *Science* 324:71–74
80. Wu G, More KL, Johnston CM, Zelenay P (2011) High-performance electrocatalysts for oxygen reduction derived from polyaniline, iron, and cobalt. *Science* 332:443–447
81. Proietti E, Jaouen F, Lefèvre M, Larouche N, Tian J, Herranz J, Dodelet JP (2011) Iron-based cathode catalyst with enhanced power density in polymer electrolyte membrane fuel cells. *Nature Comm* 2: doi:10.1038/ncomms1427

## Personal Rapid Transit and Its Development

SHANNON S. McDONALD

Southern Illinois University, Carbondale, IL, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

The Development of PRT: Personal Rapid Transit System

Key Sustainable/Energy and Planning/Architectural Future Directions

Bibliography

### Glossary

**AGT** Automated Guideway Transit – Updating the original AGT definition (US Congress, Office of Technology Assessment, 1975), Automated Guideway Transit (AGT) is defined as a class of transportation systems in which fully automated vehicles operate along dedicated guideways.

**APM** Automated People Movers – According to the General Accounting Office (1980), Automated People Movers are driverless vehicles operating on a fixed guideway. Vehicle capacities range up to 100 passengers and may be operated as single units or as trains up to 30 miles/h.

**ATRA** Advanced Transit Association

**Dual-mode** A transportation system where in one mode the vehicle operates under its own power and control, usually on existing streets and in the second mode it operates under automated control and/or external power.

**FRT** Freight Rapid Transit – Characteristics of PRT but the vehicles are designed to handle freight only.

**GRT** Group Rapid Transit – which is similar to personal rapid transit but with higher occupancy vehicles and grouping of passengers with potentially different origin–destination pairs. As noted in an early study (US Congress, 1975), the starting capacity for GRT is six passengers per car while the upper limit is around 16 or 18; there are no clear

distinctions between GRP and APM in terms of vehicle capacities.

**Network system** Interconnecting links that form the layout of transit routes and stops that constitute the total system – as opposed to a loop or corridor system

**Off-line stations** A station design where the vehicle is removed from the main line for loading or unloading allowing other vehicles to continuously flow on the main line.

**Point-to-point** The vehicle is on demand and takes the rider from his or her starting point directly to rider's destination point with no stops in between – a nonstop journey bypassing intermediate stations that relies on the use of offline stations in a network.

**PRT** Personal Rapid Transit – The definition of PRT can be a subcategory of AGT systems that offer on-demand, non-stop transportation, offline stations using small, automated vehicles on a network of dedicated guideways.

### Definition of the Subject

Personal Rapid Transit (PRT) is in the class under Automated Guideway Transit (AGT) that includes Automated People Movers, sometimes identified as a subset of APM [1, 2]. Personal Rapid Transit is a driverless automated transit technology that has a unique movement pattern for the transit rider. A rider who uses PRT will travel point to point – similar to a cab – in comparison to a “typical” transit system where the rider stays on the system through many “unnecessary” stops until reaching their final destination. PRT is a fixed system, it is transit, with predetermined stops but how you travel from Stop A to Stop B is one of the key paradigm changes for this emerging transit technology as it is not designed in its fullest implementation as a loop or corridor system. The on-demand and point-to-point approach will provide much faster travel time and more destination choices; and along with its smaller size (two to four people vehicles), operating system, and offline stations, it can attract more riders to transit due to a higher level of service.

PRT was originally designed beginning in the 1950s for the lower-density transit environments that more commonly exist in the United States. The Morgantown

system at the West Virginia University became the first driverless system placed in operation, 1972/6, and has been in continuous operation, since its opening. It is technically considered a GRT, group rapid transit system, due to the fact that the transit vehicles holds 20 people rather than the two to four people in a “true” PRT. The offline stations allow the point-to-point movement pattern to function along with computer control systems; however due to the larger vehicle, many other design possibilities of “pure” PRT did not occur [3].

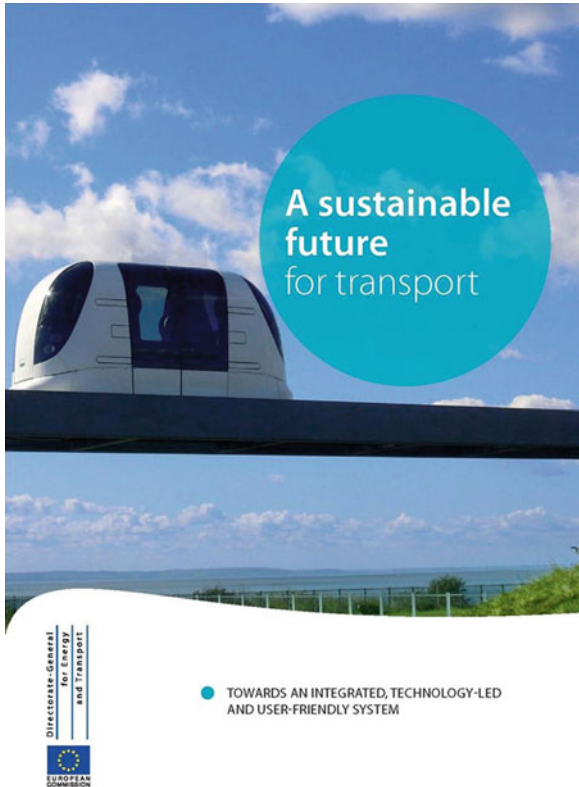
Currently, a “pure” PRT system, with small vehicles is opening at Heathrow airport connecting a long-term parking lot with a parking garage linking directly to Terminal 5. To park in this lot will be more expensive due to the direct and quick access to the airport. The computerized control systems have dramatically improved since the Morgantown system allowing the initial vision of PRT to finally be constructed and operated. The system is designed to be expanded to 50 stations, 350 vehicles, and 30 km of track connecting many different airports and local associated functions [4].

PRT not only changes the way transit functions for the user, but as well can redefine how we design our built environment, allowing for more frequent smaller transit stops encouraging walkable design and more compact development. Mazdar City, a new planned city in Abu Dhabi will rely on PRT as a major part of its transportation system. Requiring a much smaller right-of-way and lighter tracks, the infrastructure for the system is much less expensive and less visually obtrusive. Its smaller size not only reduces impact on the built environment but can also allow for alternative energy sources to power the system; such as solar and electric [5].

This system is also the perfect connector system between other forms of movement: transit, car, airplane, bus, bicycle, and walking, and for major activity centers such as downtowns, hospitals, universities, and airports, where the more traditional automated people mover system has been implemented over the years.

The reason why the PRT vision has never died and many continue to advocate for its use today is due to its potential to change our living patterns for the betterment of our society. This technology has the ability to allow for more options and flexibility of movement





**Personal Rapid Transit and Its Development. Figure 1**  
ULTra: PRT sustainable transport (Martin Lowsen of ULTra)

for all people while improving the living environment. It has been named one of the twenty-first century top 20 proven ways to save the earth by the London Times and the Sustainable Transportation Solution by the Directorate-General of Energy and Transport for the European Commission (Fig. 1).

## Introduction

The history of PRT is a confluence of new communication technology, planning visions allowing for automated control technology and transit to merge spurring innovative ideas and solutions to urban transportation providing totally new paradigms for movement and urban design. The seeds of the system started in the late 1800s and it sustained itself as an idea well into the beginning of the computer age. The early innovators and adaptors understood the power of the computer to provide new ways to manage and control these innovative transit vehicles and in turn our

movement patterns within the lower density environments of the United States.

Several individuals and companies were at the forefront of creating PRT systems. Edward O. Haltom (a contractor) Donn Fichter (a city transportation planner who believed that automated transit was the right solution for medium- to low-density populations), Monocab, TTI, Inc., Alden StaRRcar, Uniflo, Jet Rail, M.I.T, Bartells, and Dr. Jarold Kieffer each created and worked with early PRT ideas in the United States. The Urban Mass Transportation Act of 1964 and Housing and Urban Development Studies furthered the study and advanced the reality of the Morgantown system based upon the Alden StaRRcar. Numerous other innovators were involved in the early explorations, including General Motors, Raytheon, General Research Corporation, IBM, MITRE Corporation, Parsons Company, LTV Aerospace Corporation, Honeywell, Renault Engineering, Bendix, Ford Motor Company, and Otis Elevator Company along with Johns Hopkins, Ohio State, University of Minnesota, San Diego State, Battelle Columbus Laboratories, Aerospace Corporation, Jet Propulsion Lab, and Booz-Allen Applied Research [6, 7].

A series of 17 studies sponsored by Housing and Urban Development spurred the research and development of PRT. The HUD studies were summarized in a report, *Tomorrow's Transportation* by Leon Monroe Cole [8]. The two most influential studies were one by the Stanford Research Institute and the other by the General Research Corporation of Santa Barbara. The Stanford research paper was on various new concepts from moving sidewalks to PRT to dual-mode that estimated their economic benefits. The GRC study modeled alternative systems in several actual cities to compare alternative transport systems to conventional systems [9, 10].

Activity in other countries was also occurring at this time. Ed Anderson concluded that the stimulus for these explorations came from the US inventors and/or the HUD studies. The following foreign companies/countries/municipalities also advanced the work in this area: Cabtrack, CVS, Cabintaxi, Aramis, Gothenburg, Sweden and Canada. From 1968 to 1971 the not-for-profit Aerospace Corporation proved the feasibility of operating large PRT networks appropriate for urban application. The newly formed Department of Transportation led by Secretary John A. Volpe

wanted a system in operation before 1972, so a large team was put into place and the system barely opened by 1972 – a very ambitious goal for this fledgling technology; however, the Morgantown system is still functioning today. The final cost of the system was four times as planned due to many issues related to design as a result of creating a larger GRT system along with many other issues related to this. This and other early system studies are thoroughly covered in a book written by Catherine G. Burke titled: *Innovation and Public Policy: The Case of Personal Rapid Transit* [11].

However, although Morgantown, WVA, was the only system built for the public and due to its cost overruns slowed other PRT systems from being developed in the public domain; the work with and about PRT continued as the system provides so many benefits. Several test facilities built around the world proving various aspects of the concept continued and now we are seeing the benefits of this advancing technology. This concept for transit innovation has never died, and numerous attempts have been made to implement the original PRT vision of small automated point-to-point vehicles, offline stations on a dedicated network. Gayle Franzen, Chairman of the Northeastern Illinois Regional Transportation Authority initiated a new PRT program in 1990 that did not proceed.

Morgantown, WVA, is the most important historical example with the Heathrow airport system by ULTra coming into operation for the public late in 2010 leading the way into the future. Currently, several other systems and several test tracks, one in Uppsala, Sweden, a joint venture between the South Koreans and the Swedes for the Vectus system, are on line to move forward. Larry Fabian, ATRA, identifies more than 100 applications along the various spectrums of AGT technologies around the world including shuttles or circulators, major activity center circulation, and public transit, and now we can add a true PRT application to the mix [12, 13].

This article will trace the beginning PRT explorations all the way to the first implementation for public use of the complete paradigm change for transit – PRT. This has involved a complex interrelationship between innovative thinking in technology, communication systems, hardware, architecture, transportation, and urban design/planning spanning the globe.

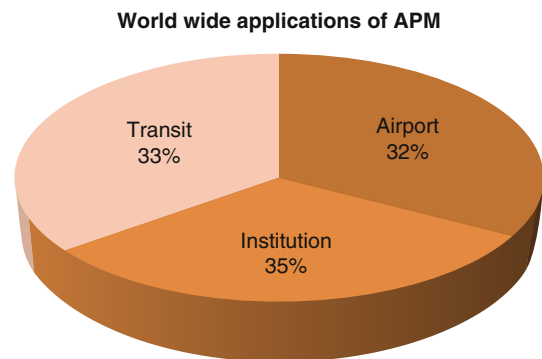
## The Development of PRT: Personal Rapid Transit System

### History/Overview/Development of Systems in the United States

The evolution of the PRT transit “idea” to its first real-world application at Heathrow airport, 2010, is to understand how a paradigm change in transit can occur, when the idea has merit, no matter how methodical and complex the route. Sending a man to the moon was easier! Only through the tenacity, belief, and hard work of many people from multiple disciplines over decades has this “vision” now occurred, so that the benefits can be “proven” – that an automated network of small personal point-to-point vehicles with offline stations can improve the built environment and as well improve transit for all people. Several papers and books have been written about this evolution; one specifically titled: *Some Early History of PRT* by Edward J. Anderson, University of Minnesota, much of the following history is a synopsis of this report [7].

A group of “transit” technologies currently known as Automated People Movers now understood as a subset of Automated Guided Transit that most of us have experienced mainly in airports around the world has been the home for the development of a true PRT system. These systems have a worldwide presence and according to Larry Fabian they are currently distributed as in Fig. 2.

In 1953, PRT was born simultaneously out of practicality and urban vision by two people who may not



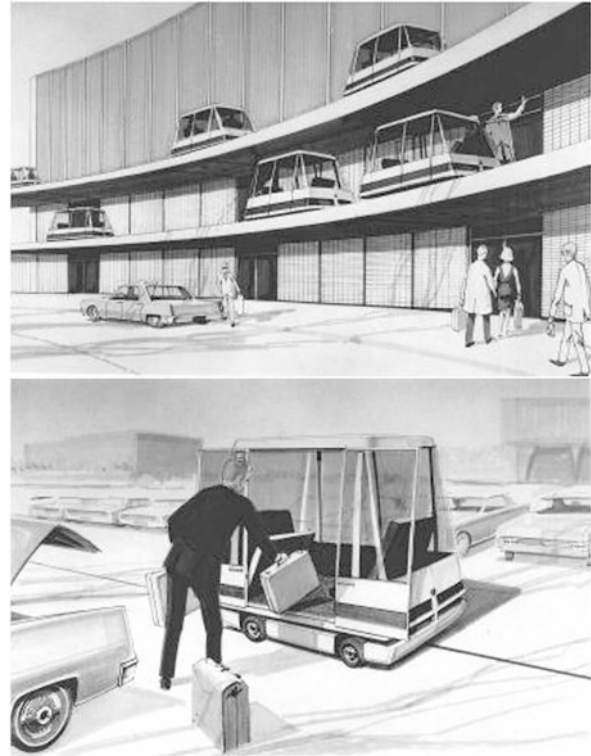
**Personal Rapid Transit and Its Development. Figure 2** Distributions of APM applications (Based on data by Fabian 2010 [1])

have initially known each other; Donn Fichter and Edward O. Haltom. Donn Fichter, a transportation graduate student in Chicago, IL, and Edward O. Haltom, a Dallas, TX, contractor who was working on a monorail system. Mr. Fichter sketched a system called Veyar and gradually developed a systems concept integrated into a city with the essential PRT ideas, although he was not a technician. He published a book in 1964 titled *Individualized Automated Transit and the City* [6]. Mr. Haltom saw a way to “improve” the monorail system that he was working on (that has been in existence since the first monorail constructed in St. Paul, Minnesota in the 1880s) by thinking of smaller guideways and smaller vehicles and invented the monocab. This system eventually became a full-scale test track in California with a totally new switching system owned by Rohr Corporation where linear induction propulsion was added to the system. The patents were purchased by Boeing and the transit system continued to be developed by UMTA (Urban Mass Transportation Administration) Advanced Group Rapid Transit until the mid-1980s.

General Motors Research Laboratories had been working for the military in the late 1950s and created a system that could become public transit, called Hovair. They formed a corporation called TTI, Inc, Transportation Technology Incorporated. A full-scale testing occurred in Detroit in 1969 and the company was eventually owned by Otis Elevator. A second test track was built in Denver. The system was a combination of air-suspension and the linear induction motor and was implemented at Duke University Medical Center. Otis also constructed several cable-drawn versions of this system.

William Alden, a graduate of the Harvard Business School invented, built, and drove a small electric vehicle that could be driven on public roads and then link to a guideway. He called it the staRRcar and it is considered the first dual-mode-system and had an on-board switching mechanism. A test track was built in Bedford, Massachusetts in 1968, the Morgantown system was based on this prototype; however a team of professionals was formed to implement the concept at Morgantown, WVA (Fig. 3).

Honeywell also in its military division through the efforts of Lloyd Berggran wanted to create urban transportation that could be competitive with the



**Personal Rapid Transit and Its Development. Figure 3**  
William Alden StaRRcar – dual-mode vehicle  
(William Alden)

automobile. He also concluded the same set of PRT characteristics. However, due to his background he created an independent “pod” with all the active power and controlling systems as part of the track. His system was called Uniflo and was designed to be in an enclosed tube. A full-scale test track was constructed by Rosemont Engineering.

Jet Rail was designed by George Adams. He designed, built, and operated a system for Braniff Airlines at Love Field in Dallas that was less expensive but similar to Monocab and proved that a lightweight guideway could be successfully constructed. Cornell Aeronautic Laboratories in the early 1960s designed Urbmobile, it was never built but the automatic control technologists were able to show that short headways could provide adequate capacity – safely. MIT in the mid-1960s produced a report call Project Metran that included the PRT ideas and influenced its continued development.

A planning director, Robert J. Bartells, and a public affairs head of a university, Dr. Jarold Kieffer, both seeking to solve real needs of transit movement for large numbers of people, independently came to the same conclusion of the PRT paradigm. Robert J. Bartells was the director of Planning for the City of Hartford, CT imagined the PRT concepts and ideas and explained and supported them as an important planning director. Dr. Jarold Kieffer, as Head of the School of Public Affairs at the University of Oregon while on a vacation at a sky resort also understood the basic concepts of PRT and has been a continual advocate for the development of the technology.

### US Government Participation

It was only through the actions of the Federal Government that the first public system was put into place both technically advancing and greatly hindering the PRT vision. Not fully understanding how totally new public-use transit paradigms need to develop if they are to have long-term success, many boasted and believed that a system could be put into place within 2 years – as we had won the race to the moon! Public-transit is a much more complex and integrated effort of many governmental agencies, organizations, and professionals.

Henry S. Reuss of Milwaukee, Wisconsin in the 1960s urged political support for the development of new transit concepts. He participated in the development of the Urban Mass Transportation Act of 1964 where a Section 6 was added focusing on research, development, and demonstration of new systems for people and goods that stated:

- ▶ The secretary shall undertake a study and prepare a program of research, development and demonstration of new systems of urban transportation that will carry people and goods within metropolitan areas speedily, safely without polluting the air, and in a manner that will contribute to sound city planning. The Program shall (1) concern itself with all aspects of new systems of urban transportation for metropolitan areas of various sizes, including technological, financial, economic, governmental and social aspects; (2) take into account the most advanced available technologies and materials; and (3) provide national leadership to efforts of states, localities, private industry, universities and foundations [12].

At this time, the US Department of Transportation did not exist; however, the Urban Mass Transportation Act established the Urban Mass Transportation Administration as a unit of the Department of Housing and Urban Development. Seventeen studies were authorized and these became known as the HUD studies and were summarized in a report titled: *Tomorrow's Transportation*. Two studies were very influential: A study by Stanford Research Institute and the General Research Corporation (GRC) of Santa Barbara. Stanford focused on new concepts from moving sidewalks to PRT to dual mode and their economic benefits while the main focus of GRC was to computer-model “advanced” systems in real cities and compares them to conventional transit.

There was a team of 17 specialists from various fields that contributed to this study identifying Boston, Houston, Hartford, and Tucson. This study strongly favored new systems due to population growth and increased use of automobiles traditional transit could not continue to meet the demand. This work published in an article, in 1969, by Scientific American titled *Systems Analysis of Urban Transportation*, became a classic for the technology and brought a national commitment to develop new technologies through the voice of Ben Alexander, the chairman of GRC [8–10].

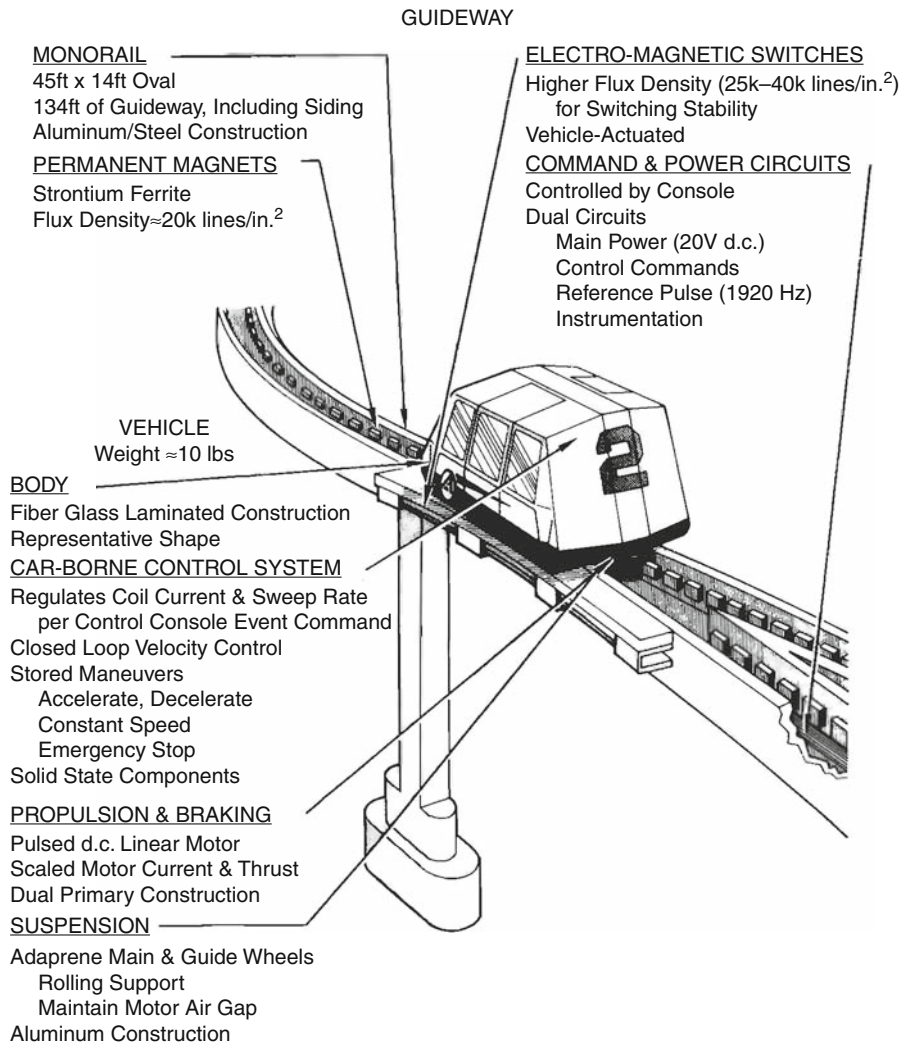
Once the HUD reports had been released The Aerospace Corporation, a not-for-profit corporation created by the United States Air Force, wanted to use aerospace technology to solve urban problems. Dr. Jack H. Irving concluded that working on high capacity PRT was a good focus for their skilled engineers. They invented new devices such as: powering the vehicles with a pair of linear pulsed direct current motors in interaction with permanent magnets in the track. Except at the switching points they used electromagnets resulting in a no-moving parts switch. The motor could be controlled by solid-state circuitry and was very efficient. They also believed that the guideway should have minimal visual impact so the vehicle was supported by two wheels in tandem. They tested this on a scale model. They were able to develop the system to a very advanced state and used computer simulations proving the feasibility of complex short headway networks. They performed analysis for Los Angeles and Tucson and released their work in

a book titled: *Fundamentals of Personal Rapid Transit, 1978* [3] (Fig. 4).

The Morgantown system was awarded and built during a very complex time for the newly formed Urban Mass Transportation Administration. Most importantly a presidential change took place bringing all new people into a process that they had no history or detailed understanding of. They wanted immediate results along with preventing the collapse of existing transit systems – two competing goals; and of course they were understaffed.

### The Development of the Morgantown GRT

The West Virginia University is located in the hilly, snowy mountain valley town of Morgantown in the Monongahela Valley of West Virginia, and its campus is spread between three different sections of the city. Buses transported students through the center of the city along with all other traffic, creating at that time extreme traffic issues. Professor Samy Elias, Head of the Industrial Engineering Department, became aware of PRT and the PRT test tracks and studies. He believed that this could be a solution for Morgantown. Due to



P

**Personal Rapid Transit and Its Development. Figure 4**  
Aerospace Corporation (Reproduced from [11]. With permission)

the support of the city, the University and the West Virginia congressional delegation money was obtained for a grant to study three systems. The Alden staRRcar was chosen as the appropriate solution for Morgantown. Due to political connections and pressures, this project was taken very seriously. So, with the 1972 presidential election in mind, the new Head of the Department of Transportation, John A. Volpe, stated that by October 1972 the president should be able to ride the system and therefore this became a federal demonstration project and a team needed to be formed. A contract was signed in December 1970 with the Jet Propulsion Laboratory in Pasadena, California, Boeing, Seattle, WA was the vehicle manufacture, Bendix, Ann Arbor, MI the control system supplier and F.R. Harris Engineering Company of Stanford, CT for the design and construction of the guideway. None of these groups had been working on PRT and there was no time for analysis or a learning curve. This of course led to a complicated and difficult outcome costing four times the original estimate as all of their expertise did not fit with the paradigm changing technology. Recently WVU was awarded a Federal Transit Authority grant that will be a part of the funds to upgrade the existing system (Fig. 5).



**Personal Rapid Transit and Its Development. Figure 5** Heathrow PRT system ( Stan Young, Kansas DOT, University of Maryland, Center for Advanced Transportation Technology)

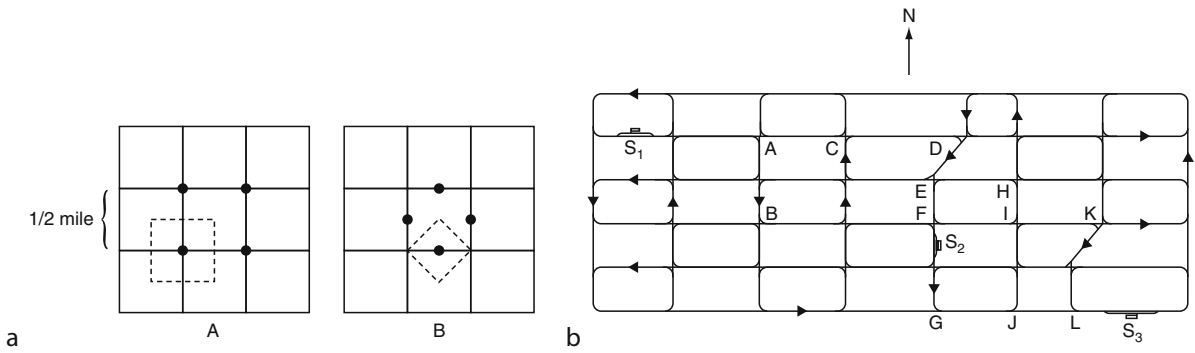
### Expositions and International Conferences

In 1972, UMTA sponsored an international exposition of four different PRT systems called Transpo72. Millions of federal dollars were set aside for the four systems to be constructed and money to be matched by their respective companies after being chosen. The companies were TTI, Monocab, Dashaveyor, and Ford. Each was to exhibit a minimum piece of guideway and one station in hopes to produce requests for capital grants. However, again due to a short time frame, work to improve the technologies took precedence and those attending the Expo did not understand how these systems could provide a better service and successfully integrate into and improve the built environment. While no city was ready to put its own funds into the development of PRT, many cities did want 100% federally funded programs to integrate the new technology. However the UMTA saw their role as one to encourage private investment [14].

In the years 1972, 1973, and 1975, three major international conferences were held by the American Society of Civil Engineers (ASCE) People Mover Committee and a volume of published proceedings were produced. However, by 1973, attendance had peaked and the organizing committee worked to develop a permanent organization. In 1976 the ATRA, Advanced Transit Association was formed. ATRA held its own conference in 1978 and more research was added to the body of knowledge. In 1988, ATRA published a broadly based technical committee report to further the importance of the PRT concept. ATRA and the ASCE – People Movers Committees – have continued their work producing published proceedings and a journal [15, 16].

### Planners' and Architects' Visions

While the visual impact of an overhead track had been identified from the beginning as an important factor in the acceptance of a PRT system, more emphasis was placed on the engineering aspects than exploring the many approaches that could support visual acceptance. Now, due to advanced technology, the first true PRT system is partially on grade and partially on elevated tracks, only where necessary within an airport, while the Mazdar City PRT system in Abu Dhabi, UAE, will be completely under the city with the vehicles on



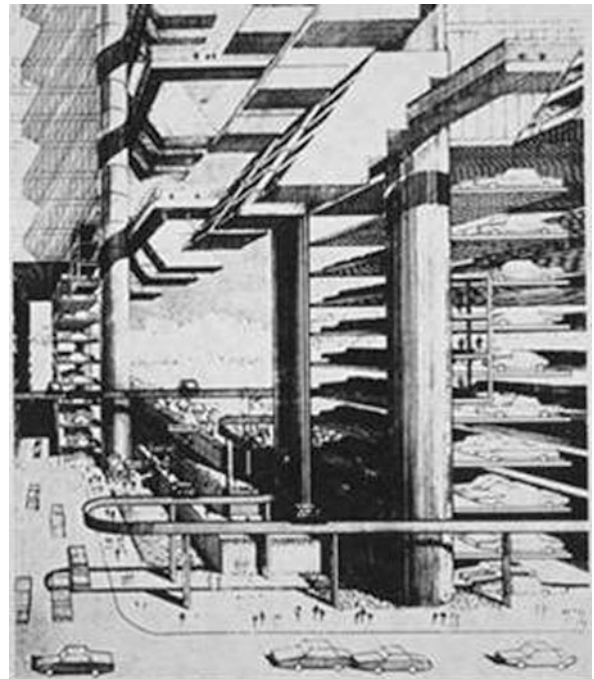
**Personal Rapid Transit and Its Development. Figure 6**

Aerospace Corporation Station placement and walking distance studies (Reproduced from [11]. With permission of Aerospace Corporation)

grade – while many lovely integrative solutions can still be explored! Some architects and architectural studies appeared during the 1970s and 1980s such as in the work of Victor Gruen, Paul Rudolph, and Ulrich Franzen.

Paul Rudolph and Ulrich Franzen were sponsored by the Ford Foundation and the American Federation of Arts producing a book in 1974 titled *Evolving City – Urban Design Proposals* [17] (Fig. 7). This publication contained amazing drawings and models for New York City including POD vehicle types and drawings that incorporated PRT into an urban setting. Victor Gruen, known as the shopping center architect, in his Fort Worth Central Business District project in 1950, envisioned “small electric vehicles” under the city to allow full access for all people to the city center. He also created the first pedestrian-only street in Kalamazoo, MI, that still exists today. The two concepts of PRT and walking environments have been connected by architects and planners since the beginning of PRT research, due to the ability of PRT to have smaller stations closer together providing a faster, more accessible, and greater choice in destination [3] (Fig. 6). PRT was also understood as a transit system of choice when connecting other forms of transportation such as automobiles and existing longer-haul transit systems (Fig. 7).

In Great Britain, a steering group for the 1963 UK Department of Transport produced with Colin Buchanan as the key author *Traffic in Towns* [18] that also became known as the Buchanan Report. He was an influential planner and this report discussed how



**Personal Rapid Transit and Its Development. Figure 7**  
Paul Rudolph and Ulrich Franzen, *Evolving City* [17]

a balance must be found as increasing automobile ownership was going to challenge the urban environment. Four key points were identified, two being: Towns should be created as environmental areas with quality of living coming first and creating a transit system that could offer an acceptable alternative. This balanced, classic, and seminal report is still referred to

today in transportation planning and it is this firm, headed by his son Malcolm Buchanan, that has led the PRT initiative at Heathrow Airport.

### History/Overview of Systems in Other Countries

In the early years of PRT, the HUD reports generated the interest in understanding and developing PRT systems around the world. Great Britain, Japan, Germany, France, Sweden, and Canada all became involved to some degree in studying and designing PRT systems.

L. R. Blake wrote an article of his own synthesis of what he saw occurring with PRT development in American and designed Autotaxi with the British-built environment in mind [19]. The company was started as a private venture, sold to Brush Electric and then was funded by the British National Research and Development Board now being called Cabtrack, a true PRT system. The system was being studied in a systematic and complete way through a step-by-step process including a contract to a large British architectural firm to integrate the system into a section of London as reported in the Architects Journal of May 1971. However, after a new election in Great Britain, the Cabtrack program was stopped. The results and reports of this process are of great value and may have contributed to the success of the Heathrow project.

The Japanese had a full-scale test facility operating in 1972 outside of Tokyo with 4.8 km of guideway and 60 vehicles. Operating with a 1-s headway and a four-passenger vehicle, a 1,000-vehicle network was simulated. Although many planning and costing studies were evaluated including one for the City of Baltimore, MD, in the United States, several design issues had not been fully considered such as the guideway, station design, and cold weather conditions.

In Germany, two companies inspired by the HUD reports were independently working on PRT concepts. They pooled their resources and thoroughly analyzed many alternatives creating a three-passenger cab supported both above and below a track. They used linear induction motors, one on each side of the vehicle and asynchronous control that was more flexible in the real world. A full-scale test facility was in operation by 1973 and was advertised in 1974 as a successful project. Called Cabintaxi, it was studied to be implemented in

Freiberg and Hagen, Germany, and was marketed in the United States. The central business district of Indianapolis tested the system. A program to build a demonstration of a 12-passenger version in Hamburg was underway, as well it was a leading competitor for the Downtown People Movers Program in Detroit, MI, sponsored by UMTA, when the economic crisis of 1980 occurred and the German government withdrew its support. Information can be found about this system as some believe its potential still exists today [20].

The French PRT system called Aramis began with Gerard Bardet in 1967 whose patents were bought by Engins Matra, and in 1970 Matra received a contract. Orly International Airport became the place for a full-scale testing, in 1974, the first phase for proof testing was complete and a contract was awarded for a public demonstration in a suburb of Paris. However the system was based on a platoon concept of grouping vehicles and was designed as a ring system around Paris so vehicle size was increased and the project eventually failed; however, the work with berthing vehicles at stations provided the facts to show that loading and unloading at stations did not slow capacity for the entire system. This project was discussed in a book titled: *Aramis or the Love of Technology* [21] (Fig. 8).

The Swedish City of Gothenburg, since their city was built on solid rock, were facing a problem of how to deal with expanding mass transportation. Excited by the British efforts they did a great deal of research on PRT systems around the world and decided that the



**Personal Rapid Transit and Its Development. Figure 8** Aramis vehicle, France (Reproduced from [11]. With permission)



technology was not ready for public deployment, but continued to be interested in the technology as it was continued to be developed. PRT studies have continued to this day in Sweden as they now move closer to an urban application.

Canada also did a comparative study of PRT, conventional highway and transit technology; however they were interested in freight as well as people movement and called PRT “Programmed Modules.” In 1973 in Ontario, the Urban Transportation Development Corporation developed a system; however due to other influences, the vehicle was designed for 40 passengers with all that that implied and met with minimal success in reference to the concept of pure PRT.

### After Morgantown

In the late 1970s money was appropriated for a study of automated transit including PRT for Indianapolis using the Cabintaxi system developed in Germany. Several different-sized passenger vehicles were tested in this comprehensive study with the result that the smallest vehicle providing the lowest total cost per passenger mile. This system had the support of business, government, and civic organizations. When this project did not proceed into a real-life application, Raymond MacDonald and J. Edward Anderson began a development program in 1981 at the University of Minnesota that was to take into account all prior research and work to develop the “perfect” PRT system, although they concluded that the aerospace PRT system was the “closest to being right” [22].

A company was formed with the assistance of the University of Minnesota and a Chicago company funded the work until 1986, when Dr. Anderson moved to Boston University. Raytheon and local Boston engineers intrigued by the Chicago-Area Regional Transportation Authority (RTA) were seeking a new approach. Two teams were organized developing designs for one to be chosen for an application. This beginning culminated in the RTA selecting the Taxi 2000 system with Raytheon to design, build, and operate the test PRT system in Rosemount, IL. In 1999, Raytheon terminated its work on PRT 2000; TAXI 2000 Corporation bought back its rights and has continued to develop its concept, now called Skyweb Express.

The PRT concept continued with 2getthere in the Netherlands implementing several successful projects, all on grade: three considered GRT and one PRT. The ParkShuttle technology was applied in the Business Park Riviam, the Airport Schiphol, and the Antibes Demonstration. The CyberCab project, a true PRT system, was tested at a public Flower Show – Floriade at Hoofddorph, Netherlands in 2002. Developed by 2getthere, all of these projects were successfully operated and Riviam continues today as a public form of transit. 2getthere is now designing the PRT system being installed as part of the new carbon neutral Masdar City further discussed below [23].

Ford appeared on the PRT scene briefly with a study of a dual-mode system called PRISM that actually is a Program for Individual Sustainable Mobility [24]. This concept had some similarities to the NEV (Neighborhood Electric Vehicle) initiative. It included very lightweight small vehicles that are on a controlled system. It is an all-electric program that would start small in a few cities eventually growing into a national network linking to high-speed rail. The reason for this program was traffic congestion and the data collected by The Texas Transportation Institute of 68 US urban areas. They understood a system like this could increase fuel efficiency, reduce emissions, and be safer as less driver error would occur.

The South Koreans had been working since 1990 on a PRT system called Vectus and in 2005 began collaborating with the Swedish to construct a test-track in Uppsala, Sweden, that is designed for winter weather application. The test facility passed safety requirements in 2007, had public access to the system in 2008 and now several studies for real projects in both Sweden and South Korea are underway. The test-track layout has an outer loop of 300 m, allowing speeds up to 12.5 m/s, and a station track of 100 m with a two-berth station. The station track is designed be long enough to allow merge operations at full speed [25].

The British Company ULtra moved forward with its demonstration project in 2002, followed by a feasibility study in 2004, for implementing PRT at Heathrow airport and is now our first “pure” PRT system in operation. A more detailed discussion follows [26].

## Personal Rapid Transit versus Existing Transportation

In understanding the paradigm-changing value of the modern pure PRT to the transportation field, a comparison of PRT to existing systems is crucial. This chart initially from Wikipedia has been adapted for this entry [27]:

Personal Rapid Transit versus Existing Transportation

<i>Similar to automobiles</i>	<ul style="list-style-type: none"> <li>• Vehicles are small – typically two to six passengers</li> <li>• Vehicles are individually used, like taxis, and shared only with the passengers of one’s choosing</li> <li>• On-demand, around-the-clock availability</li> <li>• Travel is point to point, with no intermediate stops or transfers</li> </ul>
<i>Similar to trams, buses, and monorails</i>	<ul style="list-style-type: none"> <li>• A public amenity (although not necessarily publicly owned) shared by multiple users</li> <li>• Passengers embark and disembark at discrete and fixed stations</li> <li>• Can be elevated or use air rights over existing highways – reducing land usage and congestion</li> </ul>
<i>Similar to automated people movers</i>	<ul style="list-style-type: none"> <li>• Fully automated, including vehicle control, routing, and collection of fares</li> </ul>
<i>Distinct features</i>	<ul style="list-style-type: none"> <li>• Vehicle movements are coordinated, unlike the autonomous human control of automobiles and bikes</li> <li>• Reduced local pollution (alternative energy sources such as electric power and solar)</li> <li>• Small vehicle size allows transit infrastructure to be smaller than other transit modes</li> <li>• Vehicles travel along a network of guideways where interconnection and multiple station options are standard</li> </ul>

## Current Systems in Operation or Under Construction

Currently, two GRT networks are operational, one PRT system is operational, one PRT in testing, one full-PRT network is under construction, and several more are in various stages of upgrade, planning, and design. Again the chart is adapted and updated from Wikipedia [27].

In the United States many cities, towns, and activity centers are in various stages of planning with San Jose, CA, and Ithaca, NY, currently leading the way. San Jose is working through a municipal process while Ithaca, NY is studying PRT through a NYSERDA research grant. Alameda Point, CA, Mountain View, CA, Santa Cruz, CA, that is attempting a solar powered system and Fresno, CA, November 2006, the citizens voted to spend \$36 million to establish a fund to begin a PRT effort in their downtown and is currently in the study phase, the others are in various stages of the process. Tysons Corner, VA, and Amber Glen, OR, are all also looking at PRT. In England, Bath, Cardiff, and Corby have been studied for PRT implementation. While in the United Arab Emirates, capital city Dubai, Lulu Island, Abu Dhabi, and Bawadi, Dubai, all are considering PRT systems. Several studies are beginning in India and other European countries. The technology is now being understood for its benefits to people around the world [28].

### Heathrow Airport PRT

Heathrow Airport ULTra PRT will be the first “true” PRT project for public use. It has already won a prestigious UK award for transit technology advancement and has been called one of the 20 proven ways to save the earth by the UK Times and also a Sustainable Form of Transport by the Directorate General for Energy and Transport of the European Union.

Heathrow is the main international airport in the UK and one of the busiest in the world. The airport is committed to PRT as the solution to provide connectivity within a very complicated airport system. The initial application and the pilot program connected Terminal 5 with a businessman’s parking lot. A higher

Location	Status	System	Date	Guideway	Stations/ vehicles	Notes
Morgantown, West Virginia, USA	Operational	WVU PRT	1975	13.2 km	5/73	Up to 20 passengers per vehicle, some rides not point to point during low usage periods
Schiphol Airport Amsterdam, Netherlands	Temporarily operational	Park Shuttle	1997–2004	2,000 m 2 loop network	10/4	Connection from parking to airport 24/7 operation
Rivium Business Park	Operational	Park Shuttle II	1999–today	18,000+ m	8/6	Public Transportation to Business Park and City Functions 20 passengers
Hoofddorp, Netherlands	Temporarily operational	CyberCab 2GetThere	2002	700 m	2/25	Four passengers, transport to view a flower show from a hill peak 40 m high
London Heathrow Airport, UK	Construction completed, open to public	ULTra	2010	3.8 km	3/18	Will be the world's first true commercial PRT system, initially connecting Terminal 5 with a long-term car park. If successful, BAA plans to extend it throughout the airport
Hospital Rovisco Pais, Portugal	Construction completed, under testing	Critical Move	2010			The Hospital Rovisco Pais, a center for physical rehabilitation, uses the Move between the several buildings of the hospital
Masdar City, Abu Dhabi, UAE	Under construction	2getthere	2011	Magnets in pavement, automated driving	83/2,500	Automobiles will be banned, the only powered transport will be PRT and intercity light rail
Suncheon, Republic of Korea	In planning	Vectus	2013	5 km (3.1 miles)	?/40	Will connect Suncheon to the future site of the International Gardening Festival

cost to park can be applied to the businessman as the dramatically reduced time for them from car to terminal is well worth it (Fig. 9).

The initial system is a 3.9 km (2.4 mile) single guideway connecting three stations. Having 21 vehicles to travel between the stations takes about 5 min. The system had to be constructed over a very complex site that included two rivers, seven roads, green-belt land as well as complex airport conditions. If this initial

application is successful then it will be extended to include a full network of 350 vehicles, 50 stations, and 30 km of track linking all airport private and public functions [29].

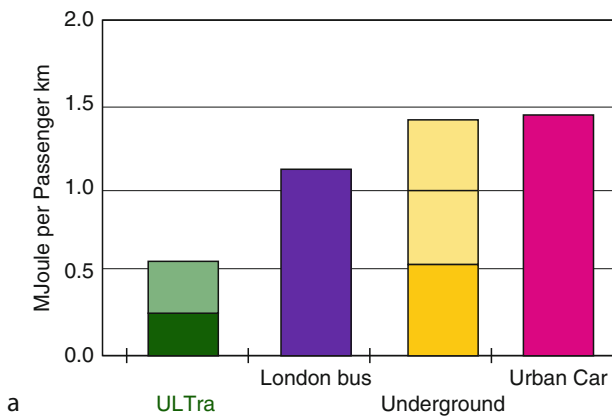
The ULTra pods can hold four adults and two children and can easily accommodate wheelchairs, bicycles and safely carry those who are physically challenged. The pods are battery controlled by a system on the pod. They will save 70% of energy compared to cars

and consume less than 50% of traditional buses [30]. The cost in comparison to light rail and transitional automatic people moves is quite less, while a traditional

bus could be less expensive, the level of service and benefits provided to the community cannot be compared (Figs. 10–12).



**Personal Rapid Transit and Its Development. Figure 9**  
Heathrow Airport PRT Phase One (Martin Lowesen of ULTra)

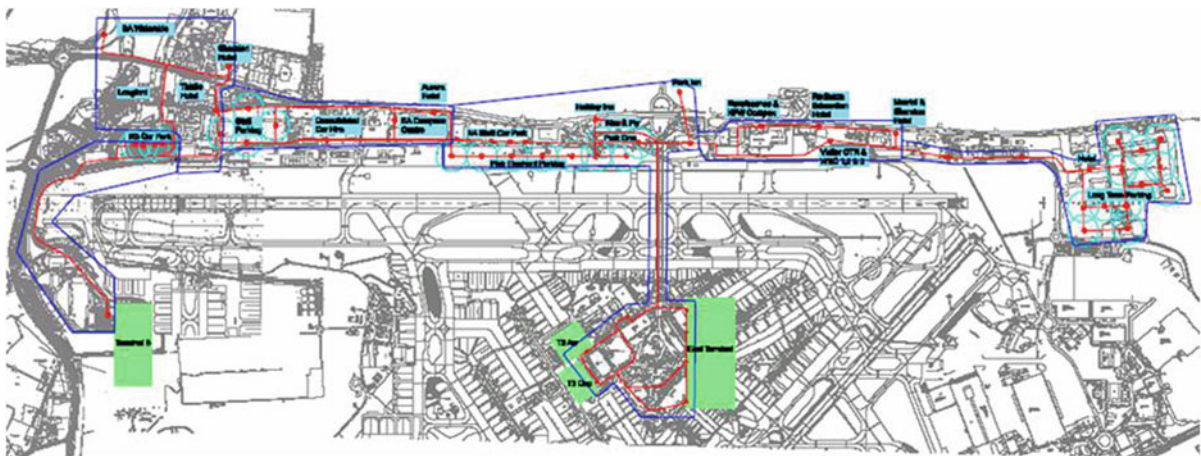


	\$m/mile
Automatic People Movers (“APM”)	30–100
Light Rail	20–40
ULTra	10–15

**Personal Rapid Transit and Its Development. Figure 10**  
ULTra: PRT sustainable transport (Martin Lowesen of ULTra)



Personal Rapid Transit and Its Development. Figure 11  
Two Heathrow Station Design (Martin Lowsen of ULTra)



Personal Rapid Transit and Its Development. Figure 12  
Heathrow expansion (Martin Lowsen of ULTra)



**Hospital Rovisco Pais, Portugal**

This is a PRT system for a hospital setting so that all within the complex have equal access to full mobility – PRT is perfect for this. It is a fully automated electric vehicle, on demand and point to point within a dedicated right-of-way. The user can have communication with the system operator while in the vehicle and can also control the position where the vehicle stops. The vehicle can accommodate eight passengers.

**Key Sustainable/Energy and Planning/ Architectural**

Being able to synergize transportation, energy, planning, and architecture to meet sustainable goals is a rare confluence of technology/people interconnections.

PRT has always been understood for its benefits in minimizing energy usage, while also providing the opportunity for the system to be designed connected to other sources of energy, new planning strategies and architectural design. Peter Calthroe, a Yale-educated architect and sustainability advocate from the 1976 calls it “The ideal transit technology: (a) stations right where you are, within walking distance, (b) no waiting.” He has been the supporter of the Alameda Point Project since 2008 especially with the discussion of an additional 60,000 housing units to the area. He advocated ULTra PRT as it would enhance carpooling and other forms of transit by allowing all of them to interconnect. PRT can also extend bicycle commutes and enhance walkable community design. At Alameda Point, PRT can assist with interconnections along

the Red and Orange lines between Fruitvale BART, Downtown/Park Street shopping, Marina Village jobs, College of Alameda and Rapid Bus transfer, Alameda Point 53-acre Sports Complex, Alameda Point National Wildlife Refuge, Bridgeside Shopping Center, Alameda Point cafes and parks, and Fruitvale Shopping Center.

### **Suncheon, Republic of Korea**

Vectus, a company founded in South Korea that teamed with the South Koreans to construct a test track in Uppsala, Sweden, is now in the planning and design phases for a project in Suncheon, S Korea. The PRT system was chosen to be constructed at the famous Suncheon Coastal Wetlands Park. It marks the first time that the governmental agency has chosen PRT over conventional transit and the environmental reasons could not be more appropriate. Including all of the characteristics that have been mentioned throughout this paper, the South Koreans chose it because it was less costly to build, takes less than half of the construction time and emits no pollutants including CO<sub>2</sub> [31].

The first phase of the system will connect the parking lot to the entrance of the wetlands and will contain 40 vehicles on a 5 km system. Mr. Dong Hee Lee, President and Chief Investment Officer of POSCO stated: “POSCO is on the forefront of the global efforts to protect the environment while improving the quality of life”. The City will host the International Garden Expo in 2013 and there are hopes that the initial system may even be expanded by this time to connect with the Central Train Station and the city’s downtown.

### **Masdar City, Abu Dhabi, UAE**

Masdar City has been under development as the world’s first zero-carbon city. This entire 6<sup>2</sup> km city is completely new and has been totally designed from scratch so that every decision is based upon creating a living place that interacts with its local environment without adversely changing it while also improving the lives of the people who live there. Recently – at the 2009 World Future Energy Summit – the cyber taxi by 2getthere was acknowledged as one of the key features in creating such an environment. They have anticipated 3,000 electric cars with 90 stops connecting light rail, parking, and linking to strategic walking points to the

city above. They will operate 24 h a day. The vehicles can travel at 40 kmph and are guided by magnets imbedded in the streets that interact with an onboard navigation system.

It is anticipated that the longest trip by PRT around the city will take no more than 10 min. The city is raised on a podium with all of the vehicles running below the pedestrian-friendly city. The PRT system is anticipated to eventually have 3,000 vehicles serving 130,000 trips/day. The PRT system will also operate freight delivery to the whole city – FRT (Freight Rapid Transit). The FRT is designed for 5,000 trips a day. Renewable energy in the form of lithium phosphate batteries allow the vehicles to be charged at the stations and will only require a 1.5 h charge for 60 km of use.

Masdar City and 2getthere have the exclusive rights to apply the FROG-technology, the supervisory planning and control system TOMS, as well as unique proprietary subsystems and components (such as the FrogBox and Magnet Ruler – MMS) and a number of related patents for Automated People Mover Systems. These are the technologies that allow this system to run on grade and they have a track record of 22+ years [32].

- ▶ The FROG (Free Ranging On Grid) technology creates intelligent vehicles that can operate in any environment. The on board FROG-box<sup>®</sup> controls the vehicle based on electronic maps (route planning). While driving, the vehicles measure distance and direction traveled by counting the number of wheel revolutions and measuring the steering angle (odometry). External reference points (magnets embedded in the road surface) are used to correct possible small inaccuracies in reference to the planned route (calibration). The reliability of the navigation system has been proven in all previous applications realized and tested successfully by TNO during the FMECA safety-procedure for the ParkShuttle Rivium application that has been operating since 1999.

### **San Jose Airport, San Jose, CA**

The City of San Jose has identified the area around the Sam Maneta airport as a place where the PRT concept can assist with linking all of the existing movement systems such as Caltrain, BART, and VTA Light Rail as well as potentially connecting to other places not directly related to the airport such as North San Jose, as

well as to new growth. They have earmarked \$4 million to conduct two simultaneous studies, one technical and the other with a focus on transportation and urban planning. Laura Stuchinsky, a sustainability officer for the San Jose Department of Transportation and many other city officials see PRT as a complement to all of these systems and one that will allow all of them to function as a complete transportation system for the airport and local area – a circulator system interconnecting all the systems. This was noted in a paper by Young, Miller, and McDonald. In September 2008, they issued the initial request that is now in the final contract negotiation phase. They expect to have an operational system by the end of 2015 (Fig. 13).

The following is an excerpt from the City of San Jose:

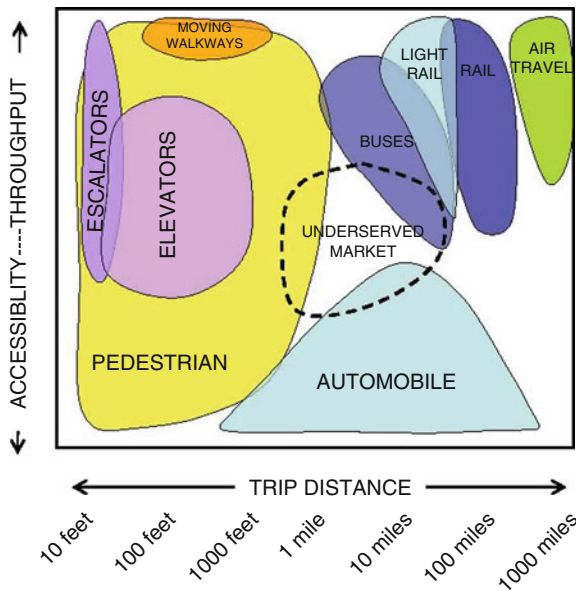
- ▶ It will start with a limited demonstration project and proceed in three phases:
  - Phase 1: Preliminary Design – A determination of feasibility as described above.

- Phase 2: Final Design – Selection of system design and construction contractor(s); completion and verification of detailed design.
- Phase 3: Construction, Integration and Testing – Physical construction of the system on-site culminating in operational certification.

All typical analysis for transportation infrastructure such as ridership forecasts, routing determinations, architectural development, civil engineering design, environmental impact assessments, and community outreach will be a part of the process. While nonconventional aspects of this project include new usage and operational models, completely new civil architectures and their integration with the as-built environment, along with new elements of the technical system – innovative designs for propulsion, control, vehicles, guideways, and vehicle management systems, for example, that may be selected, are specified and arranged in various ways to provide varying degrees of utility and capability.

New approaches and guidelines/standards may need to be developed, as will ridership forecasting and performance evaluation models. A thorough review of technical capabilities and maturity level relative to the resulting performance requirements is required as is an assessment regarding the ability of the still-small vendor base to deliver such systems.

The CITY has selected two consultants as project advisors, The Aerospace Corporation (SYSTEMS CONSULTANT) and Arup North America Limited (TRANSPORTATION CONSULTANT). In general, TRANSPORTATION CONSULTANT will address all conventional transportation infrastructure issues. SYSTEMS CONSULTANT and TRANSPORTATION CONSULTANT will together address the unconventional project issues; the TRANSPORTATION CONSULTANT focusing on the civil infrastructure-related items and the SYSTEMS CONSULTANT focusing on the technical system, as described generally above.



**Personal Rapid Transit and Its Development. Figure 13** Conceptual comparison of modes in the accessibility – throughput spectrum from Young, Miller, McDonald. Keys to innovative transport development. Presented at the 87 Annual Meeting of the Transportation Research Board, Washington, DC, 2007

**General Approach** The two consultants will work collaboratively as part of an Integrated Project Team (PROJECT TEAM) led by the CITY. They will share information and provide input necessary to each

System	Location	Active	Status	Seating Capacity (per vehicle)
Morgantown PRT (Boeing)	West Virginia	Yes	In service	8 seated plus 12 standing
ULTra (ATS Ltd)	UK	Yes	In testing for final commissioning	4
Critical Move	Portugal	Yes	In on-site testing	4
2getthere PRT	Netherlands	Yes	10 vehicles produced for Masdar City	6
Vectus PRT (POSCO)	South Korea	Yes	Full prototype, Sweden; Suncheon, Korea	4
Cabintaxi [24]	Germany	No	Completed system: 1980s approval for federal transit programs in Germany and US	3,12,18
PRT2000 (Raytheon)	USA	No	Full prototype	4
Skyweb Express (Taxi2000)	Minnesota	Yes	Partial prototype	3
MISTER	Poland	Yes	Partial prototype	5
JPods	USA	Yes	Mockup	4
SkyTaxi	Russia	Yes	Concept	1,2,4
Launchpoint Technologies SPM Maglev [25]	USA	Yes	Concept	?
Skycab AB Skycab	Sweden	Yes	Concept	?
SkyTran Unimodal	USA	Yes	Full scale prototype	3

other's work and project deliverables, integrating their workflows to achieve the CITY's project objectives [33].

### *The Basics of Several PRT Systems*

## **Future Directions**

### **High-Tech/Low-Tech Connections: Man and Machine**

As PRT becomes more available for public use, the positive connections between technology, the environment, our communities, and ourselves – the low-tech/High-tech connection will be better understood and allowed to flourish. What are the high technologies of the future and how are they transforming the ways that we travel through space, place, and time. The test is how many lives can be transformed for the better, for greater accessibility and the long-term sustainability of our planet – these technologies should be explored, developed, and expanded.

PRT with automated transit technologies will continue to evolve to become a greater part of our future built world. They have the ability to transform the way that we design and build our places for living, just as positively as the car did over 100 years ago. Following that technological innovation many peoples lives were changed and for the betterment of all. Since the late nineteenth century, world's fairs have offered futuristic visions of how people could and would live. Automated People Movers are among the visions that have become reality from them. Although most people connect monorails with world's fairs and amusement parks, there is no reason to limit advanced transit technology to such environments. For many land uses – such as airports, medical centers, and colleges – automated people movers have tremendous potential to address congestion, increase connectivity, mitigate parking shortages, and provide links to parking and PRT has the most potential within this category to really transform our lives.



Because of its human scale, PRT allows the pedestrian experience to be preserved – minus the traffic and congestion. It can also provide a more flexible link between man and machine: because it so small, quiet, and light (compared with conventional transit), PRT can be adapted to a wider array of environments. Although PRT does not eliminate the need for parking in many situations, it does offer opportunities to change the parking paradigm. In an automated parking facility, for example, PRT vehicles could easily change direction or move between levels within a small space [34].

PRT systems can provide excellent transit coverage for urban or suburban areas; such a system would be ideal, for example, in locations such as Destiny USA – the environmentally sensitive entertainment, shopping, and technology project in Syracuse, New York. PRT would be equally appropriate in more traditional urban locations, particularly as a link between parking and other land uses. As part of a larger effort to explore the transformation of office parks into transit villages, Cities21, a nonprofit research organization based in Palo Alto, California, is completing an EPA-funded study of the environmental benefits of a PRT system within the Hacienda Business Park [35].

In the 1960s, William Alden designed the StarrCar, a vehicle that could function independently, on the road, or on dedicated tracks, as part of a transit system. At the time, technology had not yet caught up to Alden's vision, although part of Alden's concept was the basis for the Morgantown PRT system. Today, however, with advanced automotive technology, Alden's vision is almost within reach. Within a system that allowed vehicles to be taken off the tracks as needed, for example, PRT vehicles may be used to pick up drivers at their parking spots and take them to their next destination. The parking facility could both provide space for traditional vehicles, and become an active participant in new approaches to transit and movement! (As automotive technologies advance, moving in the direction of "cars that drive themselves," it may be possible to drop off a car at a garage and have the car park itself as you connect to the PRT system. As PRT, automated vehicles, and other high-tech transportation options become integrated into the overall transportation system, they will affect parking garage design in ways that cannot yet be imagined [36]).

Beauty, technology, function, engineering, and civic purpose: the pieces are all in place. But determining the solutions of the future – whether urban, suburban, or rural – will mean connecting, at a deeper level, to the ways in which people relate to the world around them. It will be through these deeper connections that new solutions emerge. Horst Bredekamp, in an exploration of the phenomenon of the *Kunstammer* – the cabinets of art and curiosities that were the precursors to modern museums – notes that

- ▶ No one wants to return to the deliberate chaos of the *Kustkammer* as museums. But the boundaries between art, technology, and science are beginning to break down in a similar manner as has been demonstrated by the *Kunstammer*. In view of this fact, their lessons of visual association and thought processes which precede language systems take on a significance which might even surpass their original status. Highly technological societies are experiencing a phase of Copernican change from the dominance of language to the hegemony of images [36].

In looking toward the future of PRT, advances in new transit technology have been hindered by a lack of focus on research and development in this area. However, in fits and starts, and often with private money and due to the perseverance of innovators and visionaries, PRT is just starting on its path to its full potential, finding its way into our world of public transit linking with all existing forms of movement.

One of key area of future advancement is the control software that allows for flexibility and openness so that the complex and time-consuming tasks that will allow a large fully operational city PRT system to function safely and quickly can be applied to public transit. Systems of this type are being implemented in Automated Guided vehicles for warehouses and manufactures as well as being used successfully in airport luggage systems and emerging PRT systems.

The motivation behind PRT is to provide not just an advanced commuter transit experience but to provide a type and level of service that meets the needs of those who cannot or do not have access to the automobile, a truly new way to move and live in our modern world. It is a transit system that attempts to put people and cities first through technology while making an improvement to the quality of urban life.

The theory and reality of PRT shows this technology to be the transit technology of the future. From 2002 to 2005 the EDICT project funded by the European Union and involving 12 research organizations concluded that PRT [37]:

- Would provide future cities “a highly accessible, user-responsive, environmentally friendly transport system which offers a sustainable and economic solution”
- Could “cover its operating costs, and provide a return which could pay for most, if not all, of its capital costs”
- Would provide “a level of service which is superior to that available from conventional public transport”
- Would be “well received by the public, both public transport and car users”

The risks of being the first to implement the system are being removed and so we should see the technology move forward and provide the level of service to all users of transit that has been envisioned since its beginnings. A review of all of the literature shows many detailed specific studies about networks, walkability, systems concepts, human affect, and proposals that indicate all that is claimed about PRT is real so now according to Rachel Liu [1] all that is required are:

- ▶ multiple business models so that its many applications can find funding. As concluded in a recent PRT study (Carnegie and Hoffman 2007), AGT possesses the virtue of sustainability due to its small footprint, lower cost and lower impact on the environment. On the other hand, its small size and low-key profile have fostered a large number of applications worldwide without garnering any major headlines, which may, however, suppress its potential as a unique solution to urban circulation and congestion problems.

Along with:

on-going PRT studies reveals that the specifications of technology and assessment of costs may be relatively straightforward, but quantifying benefits associated with the implementation of a transportation project and evaluating the market conditions are complex. There are a number of analytical tools to assign a dollar value to benefits; however, some impacts such as congestion relief, safety improvements, or air quality improvements are often difficult to quantify financially. Other qualities, such as aesthetic appearance, may not even be quantifiable. Environmental and societal

impacts are often referred to as “external” effects of transportation activities since they are not directly reflected in monetary costs and benefits of project implementation. By externalizing these factors, benefit/cost analyses often do not capture the full value of beneficial impacts. However, the significance of all impacts, both positive and negative, needs to be considered in the decision-making process.

## Bibliography

### Primary Literature

1. Liu R (Rachel) (2010) Spectrum of automated guideway transit (AGT) technology and its applications. In: Kutz M (ed) Handbook of transportation engineering. McGraw-Hill, New York
2. Committee of Automated People Mover Standards (2006) Automated people mover standards, Part 3. American Society of Civil Engineers, Reston
3. Irving J (ed) (1978) Fundamentals of personal rapid transit. DC Heath, Lexington
4. Morgantown PRT: <http://transportation.wvu.edu/prt>
5. Gunter M (7 Mar 2008) Building the world's cleanest city. Fortune Magazine
6. Fichter D (1964) Individualized automated transit and the city. BH Sikes, Providence, RI
7. Anderson JE (1996) Some lessons from the history of personal rapid transit. Paper presented at the international conference on personal rapid transit (PRT) and other emerging transportation systems, Minneapolis, Minnesota, 18–20 Nov 1996. <http://advancedtransit.org/doc.aspx?id=1025>
8. Cole LM, Merritt HW (eds) (1968) Tomorrow's transportation: new systems for the urban future, US Department of Housing and Urban Development, Office of Metropolitan Development Chapters available online: “Summary,” pp. 1–5; “Recommended Future Systems,” pp. 58–77. <http://faculty.washington.edu/jbs/itrans/tomtran.htm>.
9. Henderson C et al (1968) Future urban transportation systems: descriptions, evaluations and programs, Final report, Volumes 1 and 2. Prepared for the US Department of Housing and Urban Development by the Stanford Research Institute
10. Systems Analysis of Urban Transportation Systems (1969) Scientific American 221:19–27
11. Burke C (1979) Innovation and public policy: the case of personal rapid transit. Lexington Books, D.C. Heath
12. Andreasson I (2000) Innovative transit systems, VINNOVA: Swedish Agency of Innovation Systems. [http://vinnova.se/publ/list\\_vr.htm#03](http://vinnova.se/publ/list_vr.htm#03) as well as Vectus web site <http://www.vectusprt.com/>
13. Lawson M ULTra Urban Light Transport, Advanced Transport Systems. <http://www.atstld.co.uk/>
14. United States Congress, Office of Technology Assessment (June 1975) Automated guideway transit: an assessment

- PRT and other new systems. US Government Printing Office, Washington, DC
15. Merritt HW (1993) Reflections on the New Systems Study Project, pp 35–59 in *Automated People Movers IV: Proceedings of the 4th international conference on automated people movers*, American Society of Civil Engineers, Irving, Texas, 18–20 Mar 1993. <http://faculty.washington.edu/jbs/itrans/reflec2.htm>
  16. US Congress, Office of Technology Assessment (June 1975) *Automated guideway transit: an assessment of PRT and other new systems*. Prepared for the Senate Committee on Appropriations, Transportation Subcommittee. <http://ntl.bts.gov/data/OTA/7503/7503.htm>
  17. Wolf P (1975) *Evolving city*. American Federation for the Arts, New York
  18. Reports of the Steering Group (1963) *Traffic in towns*. H.M. Stationery Office, London and reprint Penguin Books in association with HMSO
  19. Blake LR (1966) A public transport system using four-passenger, self-routing cars. *Inst. Mech. Eng. convention on guided land tTransport*, vol 181, Pt. 3 G
  20. Tadi R, Utpal D (1997) *Detroit downtown people mover: ten years after*. paper presented at the 6th international conference on automated people movers, Las Vegas, NV
  21. Latour B (1996) *Aramis or the love of technology*. Harvard University Press, Harvard
  22. MacDonald R, Anderson JE (1978) *Transit systems theory*. Lexington Books, D.C. Heath. 21st century personal rapid transit, Self-publish
  23. 2getthere: <http://www.2getthere.eu/>
  24. Proposal for Independent Sustainable Mobility (PRISM) (pdf). Ford Advanced Research Division (2003) <http://faculty.washington.edu/jbs/itrans/PRISMGPCPaper.pdf>. Accessed 3 Aug 2010
  25. <http://www.vectusprt.com/>
  26. Heathrow airport web site: [http://www.heathrowairport.com/portal/controller/dispatcher.jsp?CiID=e993760079e1d110VgnVCM10000036821c0a\\_\\_\\_\\_&ChID=10b35109350d3110VgnVCM10000036821c0a\\_\\_\\_\\_&Ct=B2C\\_CT\\_PRESS\\_RELEASE&CtID=a22889d8759a0010VgnVCM200000357e120a\\_\\_\\_\\_&ChPath=Home%5EHeathrow%5EHeathrow+press+releases](http://www.heathrowairport.com/portal/controller/dispatcher.jsp?CiID=e993760079e1d110VgnVCM10000036821c0a____&ChID=10b35109350d3110VgnVCM10000036821c0a____&Ct=B2C_CT_PRESS_RELEASE&CtID=a22889d8759a0010VgnVCM200000357e120a____&ChPath=Home%5EHeathrow%5EHeathrow+press+releases)
  27. [http://en.wikipedia.org/wiki/Personal\\_rapid\\_transit](http://en.wikipedia.org/wiki/Personal_rapid_transit)
  28. Multiple website concerning PRT <http://www.princeton.edu/~alaink/Orf467F04/NJ%20PRT%20Final%20Small.pdf>
  29. Lowson M (2004) (doc) A new approach to sustainable transport systems. 13th World Clean Air and Environmental Protection Congress London, 22–27 Aug 2004. <http://www.ultraprt.com/media/papers/>
  30. Rodgers L (2007) Are driverless pods the future. [http://news.bbc.co.uk/2/hi/uk\\_news/7148731.stm](http://news.bbc.co.uk/2/hi/uk_news/7148731.stm). Accessed Aug 2010
  31. [http://www.vectusprt.com/center/news\\_view.php](http://www.vectusprt.com/center/news_view.php)
  32. Masdar: <http://www.masdar.ae/en/home/index.aspx>
  33. San Jose: [http://www.sanjoseca.gov/transportation/SupportFiles/admin/RFI\\_Automated\\_transit\\_networkQA.pdf](http://www.sanjoseca.gov/transportation/SupportFiles/admin/RFI_Automated_transit_networkQA.pdf)
  34. Cities21: <http://www.cities21.org/cms/>
  35. McDonald S (2007) *The parking garage: design and evolution of a modern urban form*. Urban Land Institute, Washington, DC
  36. Bredekamp H (1995) *The lure of antiquity and the cult of the machine: the Kunstammer and the evolution of nature, art, and technology*. M. Wiener Publishers, Princeton, pp 113
  37. EDICT: EDICT Final Report (pdf) from [cardiff.gov.uk](http://cardiff.gov.uk)

## Books and Reviews

- Advanced Transit Association (1988) *Personal rapid transit: another option for urban transit?* *J Adv Transport* 2(38):192–314
- Anderson JE (July 2009). *An intelligent transportation network system: rationale, attributes, status, economics, benefits, and courses of study for engineers and planners*. PRT International, LLC
- Anderson J, Romig S (1974) *Personal rapid transit II*. Audio-Visual Extension, University of Minnesota, Minneapolis, MN
- Carnegie JA, Hoffman PS (2007) *Viability of personal rapid transit in New Jersey*. New Jersey Department of Transportation, New Jersey, 141p
- Institute for Transportation in Cooperation with the Advanced Transit Association (1966–present) *Advanced Transport Quarterly*, Wiley-Blackwell
- Muller P (2007) *A personal rapid Transit/Airport Automated People Mover Comparison* Proceedings of the 29th international air transport conference. American Society of Civil Engineers. <http://www.prtconsulting.com/>
- Richards B (2001) *Future transport in cities*. Spon Press, London
- Schneider JB (2011) *Innovative Transportation Systems*. This web site offers a wealth of information, including summaries and excerpts from many other sources of information. <http://faculty.washington.edu/jbs/itrans/>
- Warren R (1998) *The urban oasis: Guideways and greenways in the human environment*. McGraw-Hill, New York

## Petroleum and Oil Sands Exploration and Production

JAMES G. SPEIGHT  
CD&W Inc., Laramie, WY, USA

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Petroleum Exploration and Production

Oil Sand Exploration and Production  
 Future Directions  
 Bibliography

## Glossary

**Bitumen** A semisolid to solid hydrocarbonaceous material found filling pores and crevices of sandstone, limestone, or argillaceous sediments such as tar sand.

**Exploration** The search for petroleum using a variety of physical and spectrographic methods.

**Hot-water process** The recovery of bitumen from tar sand by use of hot water whereby the bitumen floats and the sand sinks.

**In situ conversion** Partial or complete conversion of heavy oil or tar sand bitumen in the reservoir or deposit as part of the recovery process.

**Oil mining** The recovery of petroleum using a mining method whereby an underground chamber is produced by mining and the oil is allowed or encouraged to drain into the chamber.

**Recovery** Recovery of petroleum at the surface using primary, secondary, and tertiary recovery methods.

**Tar sand mining** Recovery of tar sand by mining (digging) tar sand from the formations at or close to the surface.

## Definition of the Subject

Exploration for petroleum is an essential part of petroleum technology. Depletion of reserves is continuing at a noticeable rate and other sources of hydrocarbons are required – these include heavy oil (a type of petroleum) and tar sand bitumen.

## Introduction

Petroleum occurs in the microscopic pores of sedimentary rocks that form a reservoir – typically, reservoir rock consists of sand, sandstone, limestone, or dolomite. However, not all of the pores in a rock will contain petroleum – some will be filled with water or brine that is saturated with minerals.

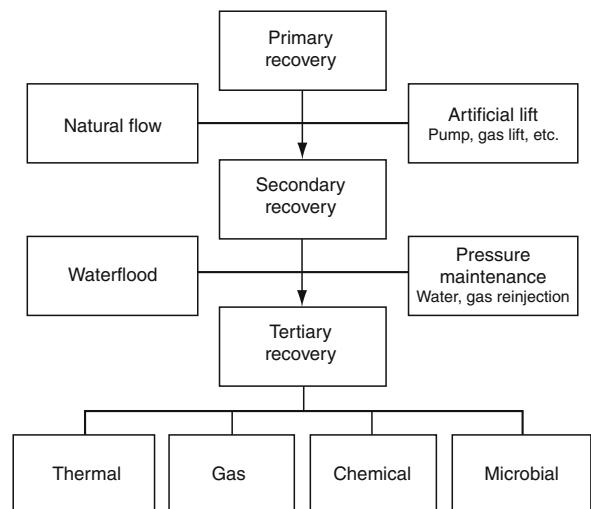
Both oil and gas have a low specific gravity relative to water and will thus float through the more porous sections of reservoir rock from their source area to the surface unless restrained by a trap. A trap is a reservoir

that is overlain and underlain by dense impermeable cap rock or a zone of very low or no porosity that restrains migrating hydrocarbon. Reservoirs vary from being quite small to covering several thousands of acres, and range in thickness from a few inches to hundreds of feet or more.

In general, petroleum is extracted by drilling wells from an appropriate surface configuration into the hydrocarbon-bearing reservoir or reservoirs. Wells are designed to contain and control all fluid flow at all times throughout drilling and producing operations. The number of wells required is dependent on a combination of technical and economic factors used to determine the most likely range of recoverable reserves relative to a range of potential investment alternatives.

There are three phases for recovering oil from reservoirs (Fig. 1):

1. *Primary recovery* occurs as wells produce because of natural energy from expansion of gas and water within the producing formation, pushing fluids into the well bore and lifting the fluids to the surface.
2. *Secondary recovery* requires energy to be applied to lift fluids to surface – this may be accomplished by injecting gas down a hole to lift fluids to the surface,



**Petroleum and Oil Sands Exploration and Production.**

**Figure 1**

Methods for oil recovery

installation of a subsurface pump, or injecting gas or water into the formation itself.

3. *Tertiary recovery* occurs when a means is required to increase fluid mobility within the reservoir – this may be accomplished by introducing additional heat into the formation to lower the viscosity (thin the oil) and improve its ability to flow to the well bore. Heat may be introduced by either (1) injecting chemicals with water (*chemical flood*, *surfactant flood*), (2) injecting steam (*steam flood*), or (3) injecting oxygen to enable the ignition and combustion of oil within the reservoir (*fire flood*).

Production rates from reservoirs depend on a number of factors, such as reservoir geometry (primarily formation thickness and reservoir continuity), reservoir pressure, reservoir depth, rock type and permeability, fluid saturations and properties, extent of fracturing, number of wells and their locations, and the ratio of the permeability of the formation to the viscosity of the oil [1, 2].

The geological variability of reservoirs means that production profiles differ from field to field. Heavy oil reservoirs can be developed to significant levels of production and maintained for a period of time by supplementing natural drive force, while gas reservoirs normally decline more rapidly.

## Petroleum Exploration and Production

### Exploration

Exploration for petroleum originated in the latter part of the nineteenth century when geologists began to map land features that were favorable for the collection of oil in a reservoir. Of particular interest to geologists were outcrops that provided evidence of alternating layers of porous and impermeable rock. The porous rock (typically a sandstone, limestone, or dolomite) provides the reservoir for the petroleum while the impermeable rock (typically clay or shale) prevents migration of the petroleum from the reservoir.

By the early part of the twentieth century, most of the areas where surface structural characteristics offered the promise of oil had been investigated and the era of subsurface exploration for oil began in the early 1920s. New geological and geophysical techniques were developed for areas where the strata were not

sufficiently exposed to permit surface mapping of the subsurface characteristics. In the 1960s, the development of geophysics provided methods for exploring below the surface of the earth.

The principles used are basically *magnetism* (*magnetometer*), *gravity* (*gravimeter*), and *sound waves* (*seismograph*). These techniques are based on the physical properties of materials that can be utilized for measurements and include those that are responsive to the methods of applied geophysics. Furthermore, the methods can be subdivided into those that focus on *gravitational properties*, *magnetic properties*, *seismic properties*, *electrical properties*, *electromagnetic properties*, *properties*, and *radioactive properties*. These geophysical methods can be subdivided into two principal groups: (1) those methods without depth control and (2) those methods having depth control.

In the first group of the measurements (those without depth control), the methods incorporate effects from both local and distant sources. For example, gravity measurements are affected by the variation in the radius of the earth with latitude. They are also affected by the elevation of the site relative to sea level, the thickness of the earth's crust, and the configuration and density of the underlying rocks, as well as by any abnormal mass variation that might be associated with a mineral deposit.

In the second group of measurements (those with depth control), seismic or electric energy is introduced into the ground and variations in transmissibility with distance are observed and interpreted in terms of geological quantities. Depths to geological horizons having marked differences in transmissibility can be computed on a quantitative basis and the physical nature of these horizons deduced.

However, geophysical exploration techniques cannot be applied indiscriminately. Knowledge of the geological parameters likely to be associated with the mineral or subsurface condition being studied is essential both in choosing the method to be applied and in interpreting the results obtained. Furthermore, not all the techniques described here may be suitable for petroleum exploration.

In petroleum exploration, terms as *geophysical borehole logging* can imply the use of one or more of the geophysical exploration techniques. This procedure involves drilling a well and using instruments to log

or make measurements at various levels in the hole by such means as *gravity (density)*, *electrical resistivity*, or *radioactivity*.

A basic rule of thumb in the upstream (or producing) sector of the oil and gas industry has been (and maybe still is in some circles of exploration technology) that the best place to find new crude oil or natural gas is near formations where it has already been found. The financial risk of doing so is far lower than that associated with drilling a rank wildcat hole in a prospective, but previously unproductive, area. On the other hand, there is a definite tradeoff between rewards for risk. The returns on drilling investment become ever leaner as more wells are drilled in a particular area because the natural distribution of oil and gas field volumes tends to be approximately log-geometric – there are only a few large fields, whereas there are a great many small ones [3].

Drilling does not end when production commences and continues after a field enters production. Extension wells must be drilled to define the boundaries of the crude oil pool. In-field wells are necessary to increase recovery rates, and service wells are used to reopen wells that have become clogged. Additionally, wells are often drilled at the same location but to different depths, to test other geological structures for the presence of crude oil.

Finally, the drilling job is complete when the drill bit penetrates the reservoir and the reservoir is evaluated to see whether the well represents the discovery of a *prospect* or whether it is a dry hole. If the hole is dry, it is plugged and abandoned.

At the stage when the prospect has been identified, reservoir evaluation is usually initiated by examining the cuttings from the well bore for evidence of hydrocarbons while the drill bit passes through a reservoir trap. The evaluation of these cuttings helps pinpoint the possible producing intervals in the well bore. At this time, a wire-line is lowered into the hole and an electric log is run to help define possible producing intervals, presence of hydrocarbons, and detailed information about the different formations throughout the well bore. Further tests (such as pressure tests, formation fluid recovery, and sidewall core analysis) can also be run on individual formations within the well bore.

If hydrocarbons are detected, the prospect becomes a *live prospect* and once the final depth has been

reached, the well is completed to allow oil to flow into the casing in a controlled manner. First, a *perforating gun* is lowered into the well to the production depth. The gun has explosive charges to create holes in the casing through which oil can flow. After the casing has been perforated, a small-diameter pipe (*tubing*) is run into the hole as a conduit for oil and gas to flow up the well and a *packer* is run down the outside of the tubing. When the packer is set at the production level, it is expanded to form a seal around the outside of the tubing. Finally, a multivalve structure (the *Christmas tree*; Fig. 2) is installed at the top of the tubing and cemented to the top of the casing. The Christmas tree allows them to control the flow of oil from the well.

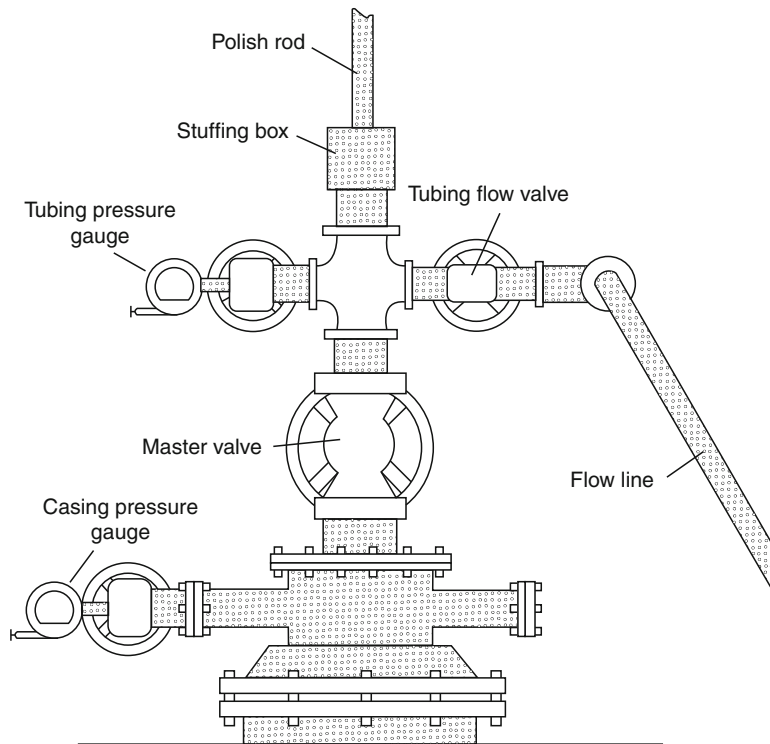
Finally, the development of an onshore shallow gas reservoir located among other established fields may be expected to incur relatively high cost and be nominally complex. A deep oil or gas reservoir located in more than 1 mile of water depth located miles away from other existing producing fields will push the limits of emerging technology at extreme costs.

Onshore developments may permit the phasing of facility investments as wells are drilled and production established to minimize economic risk. However, offshore projects may require 65% or more of the total planned investments to be made before production start-up, and impose significant economic risk.

As might be expected, the type of exploration technique employed depends upon the nature of the site. In other words, and as for many environmental operations, the recovery techniques applied to a specific site are dictated by the nature of the site and are, in fact, *site specific*. For example, in areas where little is known about the subsurface, preliminary reconnaissance techniques are necessary to identify potential reservoir systems that warrant further investigation. Techniques for reconnaissance that have been employed to make inferences about the subsurface structure include *satellite and high-altitude imagery* and *magnetic and gravity surveys*.

## Production

Recovery, as applied in the petroleum industry, is the production of oil from a reservoir. There are several methods by which this can be achieved that range from recovery due to reservoir energy (i.e., the oil flows from



Petroleum and Oil Sands Exploration and Production. Figure 2

The Christmas tree

the well hole without assistance) to enhanced recovery methods in which considerable energy must be added to the reservoir to produce the oil. However, the effect of the method on the oil and on the reservoir must be considered before application.

Generally, crude oil reservoirs sometimes exist with an overlying *gas cap*, in communication with aquifers, or both. The oil resides together with water and free gas in very small holes (pore spaces) and fractures. The size, shape, and degree of interconnection of the pores vary considerably from place to place in an individual reservoir. Below the oil layer, the sandstone is usually saturated with salt water. The oil is released from this formation by drilling a well and puncturing the limestone layer on either side of the limestone dome or fold. If the peak of the formation is tapped, only the gas is obtained. If the penetration is made too far from the center, only salt water is obtained.

Therefore, in designing a recovery project, it is a general practice to locate injection and producing wells in a regular geometric pattern so that a symmetrical and interconnected network is formed

and production can be maximized. However, the relative location of injectors and producers depends on: (1) reservoir geometry, (2) lithology, (3) reservoir depth, (4) porosity, (5) permeability, (6) continuity of reservoir rock properties, (7) magnitude and distribution of fluid saturations, and last, but certainly not least (8) fluid, i.e., oil, properties. Overall, the goal is to increase the mobility of the oil.

Once production begins, the performance of each well and reservoir is monitored and a variety of engineering techniques are used to progressively refine reserve recovery estimates over the producing life of the field. The total recoverable reserves are not known with complete certainty until the field has produced to depletion or its economic limit and abandonment.

Generally, the first stage in the extraction of crude oil is to drill a well into the underground reservoir. Often many wells (*multilateral wells*) will be drilled into the same reservoir, to ensure that the extraction rate will be economically viable. Also, some wells (*secondary wells*) may be used to pump water, steam, acids, or various gas mixtures into the reservoir to raise

or maintain the reservoir pressure, and so maintain an economic extraction rate.

*Directional drilling* is also used to reach formations and targets not directly below the penetration point or drilling from shore to locations under water [4]. A controlled deviation may also be used from a selected depth in an existing hole to attain economy in drilling costs. Various types of tools are used in directional drilling along with instruments to help orient their position and measure the degree and direction of deviation; two such tools are the *whipstock* and the *knuckle joint*. The whipstock is a gradually tapered wedge with a chisel-shaped base that prevents rotation after it has been forced into the bottom of an open hole. As the bit moves down, it is deflected by the taper about 5° from the alignment of the existing hole.

If the underground pressure in the oil reservoir is sufficient, the oil will be forced to the surface under this pressure (*primary recovery*). Natural gas (*associated natural gas*) is often present, which also supplies needed underground pressure (*primary recovery*). In this situation, it is sufficient to place an arrangement of valves (the *Christmas tree*; Fig. 2) on the well head to connect the well to a pipeline network for storage and processing.

For limestone reservoir rock, acid is pumped down the well and out the perforations. The acid dissolves channels in the limestone that lead oil into the well. For sandstone reservoir rock, a specially blended fluid containing *proppants* (sand, walnut shells, aluminum pellets) is pumped down the well and out the perforations. The pressure from this fluid makes small fractures in the sandstone that allow oil to flow into the well, while the proppants hold these fractures open. Once the oil is flowing, production equipment is set up to extract the oil from the well.

A well is always carefully controlled in its flush stage of production to prevent the potentially dangerous and wasteful *gusher*. This is actually dangerous condition, and is (hopefully) prevented by the blowout preventer and the pressure of the drilling mud. In most wells, acidizing or fracturing the well starts the oil flow.

Whatever the nature of the reservoir rock (sandstone or limestone), over the lifetime of the well the pressure will fall, and at some point, there will be

insufficient underground pressure to force the oil to the surface. *Secondary oil recovery* uses various techniques to aid in recovering oil from depleted or low-pressure reservoirs. Sometimes pumps, such as beam pumps (*horsehead pumps*) and electrical submersible pumps, are used to bring the oil to the surface. Other secondary recovery techniques increase the reservoir's pressure by water injection, natural gas reinjection, and gas lift, which injects air, carbon dioxide, or some other gas into the reservoir.

Reservoir heterogeneity, such as fractures and faults, can cause reservoirs to drain inefficiently by conventional methods. Also, highly cemented or shale zones can produce barriers to the flow of fluids in reservoirs and lead to high residual oil saturation. Reservoirs containing crude oils with low API gravity often cannot be produced efficiently without application of *enhanced oil recovery* (EOR) methods because of the high viscosity of the crude oil.

Conventional primary and secondary recovery processes are ultimately expected to produce about one third of the original oil-in-place (OOIP), although recoveries from individual reservoirs can range from less than 5% to as high as 80% v/v of the original oil-in-place. This broad range of recovery efficiency is a result of variations in the properties of the specific rock and fluids involved from reservoir to reservoir as well as the kind and level of energy that drives the oil to producing wells, where it is captured.

Conventional oil production methods may be unsuccessful because the management of the reservoir was poor or because reservoir heterogeneity has prevented the recovery of crude oil in an economical manner. In some cases, the reservoir pressure may have been depleted prematurely by poor reservoir management practices to create reservoirs with low energy and high oil saturation.

Crude oil is also produced from offshore fields, usually from steel drilling platforms set on the ocean floor. In shallow, calm waters, these may be little more than a wellhead and workspace but the larger ocean rigs include the well equipment and processing equipment as well as crew quarters. Such platforms include the *floating tension leg platform* that is secured to the sea floor by giant cables and drill ships. Such platforms can hold a steady position above a sea floor well using constant, computer-controlled adjustments. In Arctic



areas, islands may be built from dredged gravel and sand to provide platforms capable of resisting drifting ice fields.

**Primary Recovery Methods** Petroleum recovery usually starts with a formation pressure high enough to force crude oil into the well and sometimes to the surface through the tubing [5]. In this situation, it is sufficient to place the Christmas tree (Fig. 2) on the wellhead to connect the well to a pipeline network for storage and processing.

For a newly opened formation and under ideal conditions, the proportions of gas may be so high that the oil is, in fact, a solution of liquid in gas that leaves the reservoir rock so efficiently that a core sample will not show any obvious oil content. A general rough indication of this situation is a high ratio of gas to oil produced. This ratio may be zero for fields in which the rock pressure has been dissipated. The oil must be pumped out to as much as 50,000 ft<sup>3</sup> or more of gas per barrel of oil in the so-called condensate reservoirs, in which a very light crude oil (0.80 specific gravity or lighter) exists as vapor at high pressure and elevated temperature.

Crude oil moves out of the reservoir into the well by one or more of three processes. These processes are: *dissolved gas drive*, *gas cap drive*, and *water drive*. Early recognition of the type of drive involved is essential to the efficient development of an oil field.

In *dissolved gas drive (solution gas drive)* [2, 4], the propulsive force is the gas in solution in the oil, which tends to come out of solution because of the pressure release at the point of penetration of a well. Dissolved gas drive is the least efficient type of natural drive as it is difficult to control the gas-oil ratio and the bottom-hole pressure drops rapidly.

If gas overlies the oil beneath the top of the trap, it is compressed and can be utilized (*gas cap drive*) to drive the oil into wells situated at the bottom of the oil-bearing zone [2, 4]. By producing oil only from below the gas cap, it is possible to maintain a high gas-oil ratio in the reservoir until almost the very end of the life of the pool. If, however, the oil deposit is not systematically developed so that bypassing of the gas occurs, an undue proportion of oil is left behind.

Usually the gas in a gas cap (*associated natural gas*) contains methane and other hydrocarbons that may be separated out by compressing the gas. A well-known

example is *natural gasoline* that was formerly referred to as *casinghead gasoline* or *natural gas gasoline*. However at high pressures, such as those existing in the deeper fields, the density of the gas increases and the density of the oil decreases until they form a single phase in the reservoir. These are the so-called retrograde condensate pools because a decrease (instead of an increase) in pressure brings about condensation of the liquid hydrocarbons. When this reservoir fluid is brought to the surface and the condensate is removed, a large volume of residual gas remains. In many cases, this gas is recycled by compression and injection back into the reservoir, thus maintaining adequate pressure within the gas cap, and condensation in the reservoir is prevented.

The most efficient propulsive force in driving oil into a well is natural *water drive*, in which the pressure of the water forces the lighter recoverable oil out of the reservoir into the producing wells [2, 4]. In anticlinal accumulations, the structurally lowest wells around the flanks of the dome are the first to come into water. Then the oil–water contact plane moves upward until only the wells at the top of the anticline are still producing oil; eventually these also must be abandoned as the water displaces the oil. The force behind the water drive may be hydrostatic pressure, the expansion of the reservoir water, or a combination of both. Water drive is also used in certain submarine fields.

*Gravity drive* is an important factor when oil columns of several thousands of feet exist. Furthermore, the last bit of recoverable oil is produced in many pools by gravity drainage of the reservoir. Another source of energy during the early stages of withdrawal from a reservoir containing undersaturated oil is the expansion of that oil as the pressure reduction brings the oil to the bubble point (the pressure and temperature at which the gas starts to come out of solution).

The recovery efficiency for primary production is generally low when liquid expansion and solution gas evolution are the driving mechanisms. Much higher recoveries are associated with reservoirs with water and gas cap drives and with reservoirs in which gravity effectively promotes drainage of the oil from the rock pores. The overall recovery efficiency is related to how the reservoir is delineated by production wells.

For primary recovery operations, no pumping equipment is required. If the reservoir energy is not

sufficient to force the oil to the surface, then the well must be pumped. In either case, nothing is added to the reservoir to increase or maintain the reservoir energy or to sweep the oil toward the well. The rate of production from a flowing well tends to decline as the natural reservoir energy is expended. When a flowing well is no longer producing at an efficient rate, a pump is installed.

Two processes used to improve formation characteristics are *acidizing* and *fracturing*. *Acidizing* involves injecting an acid into a soluble formation, such as a carbonate, where it dissolves rock. This process enlarges the existing voids and increases permeability. *Hydraulic fracturing (fracking)* involves injecting a fluid into the formation under significant pressure that makes existing small fractures larger and creates new fractures.

Heavy oil and Tar sands (*oil sands*) have a shorter history of production and generally heavy oil reservoirs and tar sand deposits have only been subject to only one recovery technology. In the case of tar sands, primary and secondary recovery technologies, as defined for conventional oil, are not applicable because tar sand bitumen is not mobile at reservoir conditions [2, 4]. Therefore, tar sands developments generally start with a thermal recovery technology which would be considered a tertiary method or enhanced recovery method for conventional oil. However, as the development of heavy oil and tar sand technology matures, the concept of applying more than one recovery technology in a specific order is likely to also be applied to heavy oil reservoirs and tar sand deposits. In particular, in the Lloydminster area (Alberta, Canada), producers have already been investigating for several years the concept of follow-up recovery technologies once primary production is no longer economic.

**Secondary Recovery** Petroleum production is invariably accompanied by a decline in reservoir pressure and *primary recovery* comes to an end as the reservoir energy is reduced. At this stage, secondary recovery methods are applied to replace produce reservoir fluids and maintain (or increase) reservoir pressure.

*Secondary oil recovery* methods use various techniques to aid in recovering oil from depleted or low-pressure reservoirs. Sometimes pumps on the surface or submerged (electrical submersible pumps, ESPs) are

used to bring the oil to the surface. Other secondary recovery techniques increase the reservoir's pressure by water injection and gas injection, which injects air or some other gas into the reservoir. In fact, the first method recommended for improving the recovery of oil was a pressure maintenance project which involved the reinjection of natural gas, and there are indications that gas injection was utilized for this purpose before 1900 [6, 7].

The most common follow-up, or *secondary recovery*, operations usually involve the application of pumping operations or of injection of materials into a well to encourage movement and recovery of the remaining petroleum. The pump, generally known as the *horsehead pump (pump jack, nodding donkey, or sucker rod pump)*, provides mechanical lift to the fluids in the reservoir.

The up-and-down movement of the sucker rods forces the oil up the tubing to the surface. A walking beam powered by a nearby engine may supply this vertical movement, or it may be brought about through the use of a pump jack, which is connected to a central power source by means of pull rods. Depending on the size of the pump, it generally produces up to one third of a barrel of an oil-water emulsion at each stroke. The size of the pump is also determined by the depth and weight of the oil to be removed, with deeper extraction requiring more power to move the heavier lengths of polish rod.

There are also *secondary oil recovery* operations that involve the injection of water or gas into the reservoir. When water is used, the process is called a *waterflood*; when gas is used, it is called a *gas flood*. Separate wells are usually used for injection and production. The injected fluids maintain reservoir pressure or re-pressure the reservoir after primary depletion and displace a portion of the remaining crude oil to production wells.

During the withdrawal of fluids from a well, it is usual practice to maintain pressures in the reservoir at or near the original levels by pumping either gas or water into the reservoir as the hydrocarbons are withdrawn. This practice has the advantage of retarding the decline in the production of individual wells and considerably increasing the ultimate yield. It also may bring about the conservation of gas that otherwise would be wasted, and the disposal of brines that

otherwise might pollute surface and near-surface potable waters.

In the *waterflooding process*, water is injected into a reservoir to obtain additional oil recovery through movement of reservoir oil to a producing well. Generally, the selection of an appropriate flooding pattern for the reservoir depends on the quantity and location of accessible wells. Frequently, producing wells can be converted to injection wells whereas in other circumstances, it may be necessary or advantageous to drill new injection wells.

The mobility of oil is the effective permeability of the rock to the oil divided by the viscosity of the oil.

$$\lambda = k/\mu$$

where  $\lambda$  is the mobility, mD/cP,  $k$  is the effective permeability of reservoir rock to a given fluid, mD, and  $\mu$  is the fluid viscosity, cP. Thus, the mobility ratio ( $M$ ) is the mobility of the water divided by the mobility of oil:

$$M = K_{rw}\mu_o/K_{ro}\mu_w$$

where  $K_{rw}$  is the relative permeability to water,  $K_{ro}$  is the relative permeability to oil,  $\mu_o$  is the viscosity of the oil, and  $\mu_w$  is the viscosity of water.

The mobility ratio ( $M$ ) refers that  $K_o$  is the mobility of oil ahead of the front (measured at  $S_{wc}$ ) while  $K_w$  is the mobility of water at average water saturation in the water-contacted portion of the reservoir.

The mobility ratio of a waterflood will remain constant before breakthrough, but will increase after water breakthrough corresponding to the increase in water saturation and relative permeability to water in the water-contacted portion of the reservoir. Furthermore, the mobility ratio at water breakthrough is the term that is of significance in describing relative mobility ratio, i.e.,  $M < 1$  indicates a favorable displacement as oil moves faster than water and  $M = 1$  indicates a favorable displacement as both oil and water move at equal speed whereas  $M > 1$  indicates an unfavorable displacement as water moves faster than oil.

Generally, the choice of pattern (Table 1) for waterflooding must be consistent with the existing wells. The objective is to select the proper pattern that will provide the injection fluid with the maximum possible contact with the crude oil to minimize bypassing by the water.

**Petroleum and Oil Sands Exploration and Production.**

**Table 1** The ratio of injectors to producers for various well patterns

Pattern	Ratio of producing wells to injection wells	Drilling pattern required
Four spot	2	Equilateral triangle
Five spot	1	Square
Seven spot	1/2	Equilateral triangle
Inverted seven spot	2	Equilateral triangle
Nine spot	1/3	Square
Inverted nine spot	3	Square
Direct line drive	1	Rectangle
Staggered line drive	1	Offset lines of wells

In a *four-spot pattern*, the distance between all like wells is constant. Any three injection wells form an equilateral triangle with a production well at the center. The four spot may be used when the injectivity is high or the heterogeneity is minimal.

In a *five-spot pattern*, the distance between all like wells is constant. Four injection wells form a square with a production well at the center. If existing wells were drilled on square patterns, five-spot patterns (as well as nine-spot patterns) are most commonly used since they allow easy conversion to a five-spot waterflood.

In the *seven-spot pattern*, the injection wells are located at the corner of a hexagon with a production well at its center. If the reservoir characteristics yield lower than preferred injection rates, either a seven-spot (or a nine-spot) pattern should be considered because there are more injection wells per pattern than producing wells.

In the *nine-spot pattern*, the arrangement is similar to that of the five spot but with an extra injection well drilled at the middle of each side of the square. The pattern essentially contains eight injectors surrounding one producer. If existing wells were drilled on square patterns, nine-spot patterns (as well as

five-spot patterns) are most commonly used. If the reservoir characteristics yield lower injection rates than those desired, one should consider using either a nine-spot pattern (or a seven-spot pattern) where there are more injection wells per pattern than producing wells.

In the *inverted seven-spot pattern*, the arrangement is similar to the normal seven-spot pattern except where the position of the producer well was in the normal seven-spot pattern there is now an injector well. Likewise where the injector wells were in the normal seven-spot pattern, there are now producer wells. The inverted seven-spot pattern may be used when the injectivity is high or the heterogeneity is minimal.

In the *inverted nine-spot pattern*, the arrangement of the wells is similar to the normal nine-spot pattern except the position of the producer well in the normal nine-spot pattern is occupied by an injector well. Likewise where the positions of the injector wells were in the normal nine-spot, there are now producer wells. If the reservoir is fairly homogenous and the mobility ratio is unfavorable, the inverted nine-spot pattern may be promising.

In the *direct line-drive pattern*, the lines of injection and production are directly opposite to each other. If the injectivity is low or the heterogeneity is large, direct line drive is a good option. Anisotropic permeability, permeability trends, or oriented fracture systems favor line drive patterns.

In the *staggered line-drive pattern*, the wells are in lines as in the direct line, but the injectors and producers are no longer directly opposed but laterally displaced by a distance by a specified that is dependent upon the distance between wells of the same type and the distance between the lines of injector wells and producer wells. The staggered line-drive pattern is also effective for reservoirs there is anisotropic permeability or where permeability trends or oriented fracture systems.

Reservoir uniformities (and heterogeneity) also dictate the choice of pattern and mobility ratio has an important influence on pattern selection. If the ratio is unfavorable, the injectivity of an injector will exceed the productivity of a producer and water injection will supersede oil production. Hence, to balance the production with the water injection, more producers

than injectors are required. On the other hand, if the mobility ratio is favorable, the injectivity is impaired, and the pattern should have more injectors than producers.

**Enhanced Oil Recovery** Traditional primary and secondary recovery methods typically recover less than half (sometimes less than one third) of the oil only one third of the original oil-in-place. It is at some point before secondary recovery ceases to remain feasible that enhanced oil recovery methods must be applied if further oil is to be recovered.

*Enhanced oil recovery (tertiary oil recovery)* is the incremental ultimate oil that can be recovered from a petroleum reservoir over oil that can be obtained by primary and secondary recovery methods [2, 4, 8, 9]. Enhanced oil recovery methods offer prospects for ultimately producing 30–60%, or more, of the reservoir's original oil-in-place.

Enhanced oil recovery processes use *thermal, chemical, or fluid phase behavior* effects to reduce or eliminate the capillary forces that trap oil within pores, to thin the oil or otherwise improve its mobility or to alter the mobility of the displacing fluids. In some cases, the effects of gravity forces, which ordinarily cause vertical segregation of fluids of different densities, can be minimized or even used to advantage. The various processes differ considerably in complexity, the physical mechanisms responsible for oil recovery, and the amount of experience that has been derived from field application. The degree to which the enhanced oil recovery methods are applicable in the future will depend on development of improved process technology. It will also depend on improved understanding of fluid chemistry, phase behavior, and physical properties, and also on the accuracy of geology and reservoir engineering in characterizing the physical nature of individual reservoirs [10].

For taxation purposes, the Internal Revenue Service of the United States has listed the projects that qualify as enhanced oil recovery projects [11] and are therefore available for a tax credit and these projects are:

1. *Thermal recovery methods:*

Thermal methods of recovery reduce the viscosity of the crude oil by heat so that it flows more easily into the production well.

- (a) *Steam drive injection* – the continuous injection of steam into one set of wells (injection wells) or other injection source to effect oil displacement toward and production from a second set of wells (production wells).
- (b) *Cyclic steam injection* – the alternating injection of steam and production of oil with condensed steam from the same well or wells.
- (c) *In situ combustion* – the combustion of oil or fuel in the reservoir sustained by injection of air, oxygen-enriched air, oxygen, or supplemental fuel supplied from the surface to displace unburned oil toward producing wells. This process may include the concurrent, alternating, or subsequent injection of water.

Steam-based methods are the most advanced of all enhanced oil recovery methods in terms of field experience and thus have the least uncertainty in estimating performance, provided that a good reservoir description is available. Steam processes are most often applied in reservoirs containing heavy crude oil, usually in place of rather than following secondary or primary methods. Commercial application of steam processes has been underway since the early 1960s.

## 2. Gas flood recovery methods:

- (a) *Miscible fluid displacement* – the injection of gas (e.g., natural gas, enriched natural gas, a liquefied petroleum slug driven by natural gas, carbon dioxide, nitrogen, or flue gas) or alcohol into the reservoir at pressure levels such that the gas or alcohol and reservoir oil are miscible.
- (b) *Carbon dioxide-augmented waterflooding* – the injection of carbonated water, or water and carbon dioxide, to increase waterflood efficiency.
- (c) *Immiscible carbon dioxide displacement* – the injection of carbon dioxide into an oil reservoir to effect oil displacement under conditions in which miscibility with reservoir oil is not obtained; this process may include the concurrent, alternating, or subsequent injection of water.
- (d) *Immiscible nonhydrocarbon gas displacement* – the injection of nonhydrocarbon gas (e.g., nitrogen) into an oil reservoir, under

conditions in which miscibility with reservoir oil is not obtained, to obtain a chemical or physical reaction (other than pressure) between the oil and the injected gas or between the oil and other reservoir fluids; this process may include the concurrent, alternating, or subsequent injection of water.

## 3. Chemical flood recovery methods:

Three enhanced oil recovery processes involve the use of chemicals – surfactant/polymer, polymer, and alkaline flooding [12]. However, each reservoir has unique fluid and rock properties, and specific chemical systems must be designed for each individual application. The chemicals used, their concentrations in the slugs, and the slug sizes depend upon the specific properties of the fluids and the rocks involved and upon economic considerations.

- (a) *Surfactant flooding* is a multiple-slug process involving the addition of surface-active chemicals to water [13]. These chemicals reduce the capillary forces that trap the oil in the pores of the rock. The surfactant slug displaces the majority of the oil from the reservoir volume contacted, forming a flowing oil–water bank that is propagated ahead of the surfactant slug. The principal factors that influence the surfactant slug design are interfacial properties, slug mobility in relation to the mobility of the oil–water bank, the persistence of acceptable slug properties and slug integrity in the reservoir, and cost.
- (b) *Microemulsion flooding* also known as *surfactant-polymer flooding* involves injection of a surfactant system (e.g., a surfactant, hydrocarbon, cosurfactant, electrolyte, and water) to enhance the displacement of oil toward producing wells; and [2] *caustic flooding* – the injection of water that has been made chemically basic by the addition of alkali metal hydroxides, silicates, or other chemicals.
- (c) *Polymer-augmented waterflooding* – the injection of polymeric additives with water to improve the areal and vertical sweep efficiency of the reservoir by increasing the viscosity and decreasing the mobility of the water injected; polymer-augmented waterflooding does not include the injection of polymers for the

purpose of modifying the injection profile of the wellbore or the relative permeability of various layers of the reservoir, rather than modifying the water-oil mobility ratio.

Certain types of reservoirs, such as those with very viscous crude oils and some low-permeability carbonate (limestone, dolomite, or chert) reservoirs, respond poorly to conventional secondary recovery techniques. The viscosity (or the API gravity) of petroleum is an important factor that must be taken into account when heavy oil is recovered from a reservoir.

In these reservoirs, it is desirable to initiate *enhanced oil recovery* operations as early as possible. This may mean considerably abbreviating conventional secondary recovery operations or bypassing them altogether.

*Thermal methods* for oil recovery have found most use when the oil in the reservoir has a high viscosity. For example, heavy oil is usually highly viscous (hence the use of the adjective *heavy*), with a viscosity ranging from approximately 100 cP to several million centipoises at the reservoir conditions. In addition, oil viscosity is also a function of temperature and API gravity [2, 14]. Thus, for heavy crude oil samples with API gravity ranging from 4 to 21°API (1.04–0.928 kg/m<sup>3</sup>):

$$\log \log(\mu\sigma + \alpha) = A - B \log(T + 460)$$

In this equation,  $\mu\sigma$  is oil viscosity in cP,  $T$  is temperature in °F,  $A$  and  $B$  are constants, and  $\alpha$  is an empirical factor used to achieve a straight-line correlation at low viscosity. This equation is usually used to correlate kinematic viscosity in centistokes, in which case an  $\alpha$  of 0.6–0.8 is suggested (dynamic viscosity in cP equals kinematic viscosity in cSt times density in g/mL).

An alternative equation for correlating viscosity data (where  $a$  and  $b$  are constants, and  $T^*$  is the absolute temperature) is:

$$\mu = ae^{b/T^*}$$

*Thermal-enhanced oil recovery processes* add heat to the reservoir to reduce oil viscosity and/or to vaporize the oil. In both instances, the oil is made more mobile so that it can be more effectively driven to producing wells. In addition to adding heat, these processes

provide a driving force (pressure) to move oil to producing wells.

*Steam drive injection (steam injection)* has been commercially applied since the early 1960s. The process occurs in two steps: (1) steam stimulation of production wells, that is, direct steam stimulation and (2) steam drive by steam injection to increase production from other wells (indirect steam stimulation).

When there is some natural reservoir energy, steam stimulation normally precedes steam drive. In steam stimulation, heat is applied to the reservoir by the injection of high-quality steam into the production well. This cyclic process, also called *huff and puff* or *steam soak*, uses the same well for both injection and production. The period of steam injection is followed by production of reduced viscosity oil and condensed steam (water). One mechanism that aids production of the oil is the flashing of hot water (originally condensed from steam injected under high pressure) back to steam as pressure is lowered when a well is put back on production.

*Cyclic steam injection* is the alternating injection of steam and production of oil with condensed steam from the same well or wells. Thus, steam generated at surface is injected in a well and the same well is subsequently put back on production.

A cyclic steam injection process includes three stages. The first stage is injection, during which a measured amount of steam is introduced into the reservoir. The second stage (*the soak period*) requires that the well be shut in for a period of time (usually several days) to allow uniform heat distribution to reduce the viscosity of the oil (alternatively, to raise the reservoir temperature above the pour point of the oil). Finally, during the third stage, the now-mobile oil is produced through the same well. The cycle is repeated until the flow of oil diminishes to a point of no returns.

The high gas mobility may limit recovery through its adverse effect on the sweep efficiency of the burning front. Because of the density contrast between air and reservoir liquids, the burning front tends to override the reservoir liquids. To date, combustion has been most effective for the recovery of viscous oils in moderately thick reservoirs in which reservoir dip and continuity provide effective gravity drainage or operational factors permit close well spacing.

Using combustion to stimulate oil production is regarded as attractive for deep reservoirs [15] and, in contrast to steam injection, usually involves no loss of heat. The duration of the combustion may be short (<30 days) or more prolonged (approximately 90 days), depending upon requirements. In addition, backflow of the oil through the hot zone must be prevented or coking occurs.

Both forward and reverse combustion methods have been used with some degree of success when applied to tar sand deposits. The forward-combustion process has been applied to the Orinoco deposits [16] and in the Kentucky sands [15]. The reverse combustion process has been applied to the Orinoco deposit [17] and the Athabasca [2, 4]. In tests such as these, it is essential to control the airflow and to mitigate the potential for spontaneous ignition [17]. A modified combustion approach has been applied to the Athabasca deposit [2, 4]. The technique involved a heat-up phase and a production (or blowdown phase) followed by a displacement phase using a fireflood-waterflood (COFCAW) process.

## Oil Sand Exploration and Production

Heavy oil and bitumen (the component of interest in tar sand) are often defined (loosely and incorrectly) in terms of API gravity. A more appropriate definition of bitumen, which sets it aside from heavy oil and conventional petroleum, is based on the definition offered by the US government as the *extremely viscous hydrocarbon which is not recoverable in its natural state by conventional oil well production methods including currently used enhanced recovery techniques* [2, 4].

By inference, conventional petroleum and heavy oil (recoverable by *conventional oil well production methods including currently used enhanced recovery techniques*) are different to tar sand bitumen. Be that as it may, in some stage of production, conventional petroleum (in the later stages of recovery) and heavy oil (in the earlier stages of recovery) may require the application of enhanced oil recovery methods for recovery.

## Oil Mining

Oil mining includes recovery of oil and/or heavy oil by drainage from reservoir beds to mine shafts or other openings driven into the rock or by drainage from the

reservoir rock into mine openings driven outside the reservoir but connected with it by boreholes or mine wells.

Oil mining methods should be applied in reservoirs that have significant residual oil saturation and have reservoir or fluid properties that make production by conventional methods inefficient or impossible. The high well density in improved oil mining usually compensates for the inefficient production caused by reservoir heterogeneity.

However, close well spacing can also magnify the deleterious effects of reservoir heterogeneity. If a high-permeability streak exists with a lateral extent that is less than the inter-well spacing of conventional wells but is comparable to that of improved oil mining, the channeling is more unfavorable for the improved oil mining method.

## Tar Sand Mining

The bitumen occurring in tar sand deposits poses a major recovery problem. The material is notoriously immobile at formation temperatures and must therefore require some stimulation (usually by thermal means) in order to ensure recovery. Alternately, proposals have been noted which advocate bitumen recovery by solvent flooding or by the use of emulsifiers. There is no doubt that with time, one or more of these functions may come to fruition, but for the present, the two commercial operations rely on the mining technique.

The alternative to in situ processing is to mine tar sand, transport the mined material to a processing plant, extract the bitumen, and dispose of the waste sand. Such a procedure is often referred to as *oil mining*. This is the term applied to the surface or subsurface excavation of petroleum-bearing formations for subsequent removal of the heavy oil or bitumen by washing, flotation, or retorting treatments.

The tar sand mining method of recovery has received considerable attention since it was chosen as the technique of preference for the only two commercial bitumen recovery plants in operation in North America. In situ processes have been tested many times in the United States, Canada, and other parts of the world and are ready for commercialization. There are also conceptual schemes that are a combination of

both mining (aboveground recovery) and in situ (non-mining recovery) methods.

Engineering a successful oil mining project must address a number of items because there must be sufficient recoverable resources, the project must be conducted safely, and the project should be engineered to maximize recovery within economic limits. The use of a reliable screening technique is necessary to locate viable candidates. Once the candidate is defined, this should be followed by an exhaustive literature search covering the local geology, drilling, production, completion, and secondary and tertiary recovery operations.

The reservoir properties, which can affect the efficiency of heavy oil or bitumen production by mining technology, can be grouped into three classes:

1. *Primary properties*, i.e., those properties that have an influence on the fluid flow and fluid storage properties and include rock and fluid properties, such as porosity, permeability, wettability, crude oil viscosity, and pour point
2. *Secondary properties*, i.e., those properties that significantly influence the primary properties, including pore size distribution, clay type, and content
3. *Tertiary properties*, i.e., those other properties that mainly influence oil production operation (fracture breakdown pressure, hardness, and thermal properties) and the mining operations (e.g., temperature, subsidence potential, and fault distribution)

There are also important rock mechanical parameters of the formation in which a tunnel is to be mined and from where all oil mining operations will be conducted. These properties are mostly related to the mining aspects of the operations, and not all are of equal importance in their influence on the mining technology. Their relative importance also depends on the individual reservoir.

Surface mining is the mining method that is currently being used by Suncor Energy and Syncrude Canada Limited to recover tar sand from the ground. Surface mining can be used in mineable tar sand areas which lie under 250 ft or less of overburden material. Less than 10% of the Athabasca Oil Sands deposit can be mined using the surface mining technique, as the other 90% of the deposit has more than 250 ft of

overburden. This other 90% will have to be mined using different mining techniques.

There are two methods of mining currently in use in the Athabasca Oil Sands. Suncor Energy uses the truck and shovel method of mining whereas Syncrude uses the truck and shovel method of mining, as well as draglines and bucket-wheel reclaimers. These enormous draglines and bucket-wheels are being phased out and soon will be completely replaced with large trucks and shovels. The shovel scoops up the tar sand and dumps it into a heavy hauler truck. The heavy hauler truck takes the tar sand to a conveyor belt that transports the tar sand from the mine to the extraction plant. Presently, there are extensive conveyor belt systems that transport the mined tar sand from the recovery site to the extraction plant. With the development of new technologies, these conveyors are being phased out and replaced with hydrotransport technology.

Hydrotransport is a combination of ore transport and preliminary extraction. After the bituminous sands have been recovered using the truck and shovel method, it is mixed with water and caustic soda to form a slurry and is pumped along a pipeline to the extraction plant. The extraction process thus begins with the mixing of the water and agitation needed to initiate bitumen separation from the sand and clay.

Mine spoils need to be disposed of in a manner that assures physical stabilization. This means appropriate slope stability for the pile against not only gravity but also earthquake forces. Since return of the spoils to the mine excavations is seldom economical, the spoil pile must be designed as a permanent structure whose outline blends into the landscape. Straight, even lines in the pile must be avoided.

Underground mining options have also been proposed but for the moment have not been developed because of the fear of collapse of the formation onto any operation/equipment. This particular option should not, however, be rejected out-of-hand because a novel aspect or the requirements of the developer (which remove the accompanying dangers) may make such an option acceptable.

The tar sand recovered by mining is sent to the processing plant for separation of the bitumen from the sand prior to upgrading.



## The Hot-Water Process

The *hot-water process* is, to date, the only successful commercial process to be applied to bitumen recovery from mined tar sands in North America [18–22]. Many process options have been tested with varying degrees of success, and one of these options may even supersede the hot-water process.

The process utilizes (1) the film of water coats most of the mineral matter, which permits extraction by the *hot-water process*, (2) the linear and the nonlinear variation of bitumen density and water density, respectively, with temperature so that the bitumen that is heavier than water at room temperature becomes lighter than water at 80°C (180°F), and (3) natural surface-active materials (surfactants) in the tar sand which also contribute to freeing the bitumen from the sand.

In the process, the tar sand is introduced into a *conditioning* drum where the sand is heated and mixed with water to encourage agglomeration of the oil particles. Conditioning is carried out in a slowly rotating drum that contains a steam-sparging system for temperature control as well as mixing devices to assist in lump size reduction and a size ejector at the outlet end. The tar sand lumps are reduced in size by ablation and mixing action. The conditioned *pulp* has the following characteristics: (1) solids 60–85% and (2) pH 7.5–8.5.

Lumps of as-mined tar sand are reduced in size by ablation, and the conditioned pulp is screened through a double-layer vibrating screen. Water is then added to the screened material (to achieve more beneficial pumping conditions), and the pulp enters the separation cell through a central feed well and distributor. The bulk of the sand settles in the cell and is removed from the bottom as tailing, but the majority of the bitumen floats to the surface and is removed as froth. A middlings stream (mostly water with suspended fines and some bitumen) is withdrawn from approximately midway up the side of the cell wall. Part of the middlings is recycled to dilute the conditioning-drum effluent for pumping. Clays do not settle readily and generally accumulate in the middlings layer. High concentrations of clays increase the viscosity and can prevent normal operation in the separation cell.

The separation cell acts like two settlers – one on top of the other – and in the lower settler, the sand

settles down, whereas in the upper settler, the bitumen floats. The bulk of the sand in the feed is removed from the bottom of the separation cell as tailings. A large portion of the feed bitumen floats to the surface of the separation cell and is removed as bituminous froth. A middlings stream consists mostly of water with some suspended fine minerals and bitumen particles, and a portion of the middlings may be returned for mixing with the conditioning-drum effluent in order to dilute the separation-cell feed for pumping. The remainder of the middlings is withdrawn from the separation cell to be rejected after processing in the scavenger cells.

The combined froth from the separation cell and scavenging operation contains an average of about 10% by weight mineral material and up to 40% by weight water. The dewatering and demineralizing is accomplished in two stages of centrifuging; in the first stage, the coarser mineral material is removed but much of the water remains. The feed then passes through a filter to remove any additional large-size mineral matter that would plug up the nozzles of the second stage centrifuges.

In the scavenging cell, froth flotation with air is usually employed to recover more bitumen. The scavenger froth is combined with the separation-cell froth to be further treated and upgraded to synthetic crude oil. Tailings from the scavenger cell join the separation-cell tailings stream and go to waste.

The bituminous froth from the hot-water process may be mixed with a hydrocarbon diluent, e.g., coker naphtha, and centrifuged. The Suncor process employs a two-stage centrifuging operation, and each stage consists of multiple centrifuges of conventional design installed in parallel. The bitumen product contains 1% by weight to 2% by weight mineral (dry bitumen basis) and 5% by weight to 15% by weight water (wet diluted basis). Syncrude also utilizes a centrifuge system with naphtha diluent.

One of the major problems that arises from the hot-water process is the disposal and control of the tailings. The fact is that each ton of tar sand in place has a volume of about 16 ft<sup>3</sup>, which will generate about 22 ft<sup>3</sup> of tailings giving a volume gain on the order of 40%. If the mine produces about 200,000 t of tar sand per day, the volume expansion represents a considerable solids disposal problem. Tailings from

the process consist of about 49–50% by weight of sand, 1% by weight of bitumen, and about 50% by weight of water. The average particle size of the sand is about 200  $\mu\text{m}$ , and it is a suitable material for dike building. Accordingly, Suncor used this material to build the sand dike, but for fine sand, the sand must be well compacted.

Environmental regulations in Canada or the United States will not allow the discharge of tailings streams into (1) the river; (2) on to the surface; or (3) on to any area where contamination of groundwater domains or the river may be contaminated. The tailings streams is essentially high in clays and contains some bitumen, hence the current need for tailings ponds, where some settling of the clay occurs. In addition, an approach to acceptable reclamation of the tailings ponds will have to be accommodated at the time of site abandonment.

The structure of the dike may be stabilized on the upstream side by beaching. This gives a shallow slope but consumes sand during the season when it is impossible to build the dike. In remote areas such as the Fort McMurray (Alberta) site, the dike can only be built in above-freezing weather because (1) frozen water in the pores of the dike will create an unstable layer and (2) the vapor emanating from the water creates a fog, which can create a work hazard. The slope of the tailings dike is about 2.5:1 depending on the amount of fines in the material. It may be possible to build with 2:1 slopes with coarser material, but steeper slopes must be stabilized quickly by beaching. After discharge from the hot-water separation system, it is preferable that attempts be made to separate the sand, sludge, and water, hence, the tailings pond. The sand is used to build dikes and the runoff that contains the silt, clay, and water collects in the pond. Silt and some clay settle out to form sludge, and some of the water is recycled to the plant.

In summary, the hot-water separation process involves extremely complicated surface chemistry with interfaces among various combinations of solids (including both silica sand and aluminosilicate clays), water, bitumen, and air. The control of pH is critical with the preferred range being 8.0–8.5, which is achievable by use of any of the monovalent bases. Polyvalent cations must be excluded because they tend to flocculate the clays and thus raise the viscosity of the middlings in the separation cell.

## Other Processes

The issues arising from bitumen mining and bitumen recovery may be alleviated somewhat by the development of process options that require considerably less water in the sand/bitumen separation step. Such an option would allow a more gradual removal of the tailings ponds.

A *cold-water process* for bitumen separation from mined tar sand has also been recommended [23, 24]. The process uses a combination of cold water and solvent, and the first step usually involves disintegration of the tar sand charge that is mixed with water, diluent, and reagents. The diluent may be a petroleum distillate fraction such as aromatic naphtha or kerosene and is added in approximately a 1:1 weight ratio to the bitumen in the feed. The pH is maintained at 9–9.5 by the addition of wetting agents and approximately 0.77 kg of soda ash per ton of tar sand. The effluent is mixed with more water, and in a raked classifier, the sand is settled from the bulk of the remaining mixture. The water and oil overflow the classifier and are passed to thickeners where the oil is concentrated. Clay in the tar sand feed has a distinct effect on the process; it forms emulsions that are hard to break and are wasted with the underflow from the thickeners.

The *sand-reduction process* is a cold-water process without solvent. In the first step, the tar sand feedstock is mixed with water at approximately 20°C (68°F) in a screw conveyor in a ratio of 0.75–3 t per ton of tar sand (the lower range is preferred). The mixed pulp from the screw conveyor is discharged into a rotary-drum screen, which is submerged in a water-filled settling vessel. The bitumen forms agglomerates that are retained by an 840- $\mu\text{m}$  (20-mesh) screen. These agglomerates settle and are withdrawn as oil product. The sand readily passes through the 840  $\mu\text{m}$  (20 mesh) screen and is withdrawn as waste stream. The process is called sand reduction because its objective is the removal of sand from the tar sand to provide a feed suitable for a fluid coking process; ca 80% of sand is removed. Nominal composition of the oil product is 58% by weight (bitumen), 27% by weight mineral matter, and 15% by weight water.

The *spherical agglomeration process* resembles the sand-reduction process. Water is added to tar sands and the mixture is ball-milled. The bitumen forms

dense agglomerates of 75% by weight to 87% by weight bitumen, 12% by weight to 25% by weight sand, and 1% by weight to 5% by weight water.

An *oleophilic sieve process* [25, 26] offers the potential for reducing tailings pond size because of a reduction in the water requirements. The process is based on the concept that when a mixture of an oil phase and an aqueous phase is passed through a sieve made from oleophilic materials, the aqueous phase and any hydrophilic solids pass through the sieve but the oil adheres to the sieve surface on contact. The sieve is in the form of a moving conveyor, the oil is captured in a recovery zone, and recovery efficiency is high.

An anhydrous *solvent extraction process* for bitumen recovery has been attempted and usually involves the use of a low-boiling hydrocarbon. The process generally involves up to four steps. In the mixer step, fresh tar sand is mixed with recycle solvent that contains some bitumen and small amounts of water and mineral and the solvent-to-bitumen weight ratio is adjusted to approximately 0.5. The drain step consists of a three-stage countercurrent wash. Settling and draining time is approximately 30 min for each stage. After each extraction step, a bed of sand is formed and the extract is drained through the bed until the interstitial pore volume of the bed is emptied. The last two steps of the process are devoted to solvent recovery solvent recovery from the bitumen and from the solids, which holds the key to the economic success the process.

Another aboveground method of separating bitumen from mined tar sand involves *direct heating of the tar sand* without previous separation of the bitumen [27]. Thus, the bitumen is not recovered as such but is an upgraded overhead product. In the process, the sand is crushed and introduced into a vessel, where it is contacted with either hot (spent) sand or with hot product gases that furnish part of the heat required for cracking and volatilization. The volatile products are passed out of the vessel and are separated into gases and (condensed) liquids. The coke that is formed as a result of the thermal decomposition of the bitumen remains on the sand, which is then transferred to a vessel for coke removal by burning in air. The hot flue gases can be used either to heat incoming tar sand or as refinery fuel. As expected, processes of this type yield an upgraded product but

require various arrangements of pneumatic and mechanical equipment for solids movement around the refinery.

In *improved mining*, directional (horizontal or slant) wells are drilled into the reservoir from a mine in an underlying formation to drain oil by pressured depletion and gravity drainage. In the process of gravity drainage extraction of liquid crude oil, the wells are completed so that only the forces acting within the reservoir are used. A large number of closely spaced wells can be drilled into a reservoir from an underlying tunnel more economically than the same number of wells from the surface. In addition, only one pumping system is required in underground drainage, whereas at the surface, each well must have a pumping system. The objective of using a large number of wells is to produce each well slowly so that the gas–oil and water–oil interfaces move toward each other efficiently. By maintaining the reservoir pressures because of forces acting on the reservoir, it is then assured that the oil production is provided by the internal forces due to gravity (the buoyancy effect) and capillary effects.

Large vertical shafts sunk from the surface are generally the means through which underground openings can be excavated. These shafts are one means of access to offer an outlet for removal of excavated rock, provide sufficient opening for equipment, provide ventilation, and allow the removal of oil and gas products during later production. These requirements plus geological conditions and oil reservoir dimensions determine the shaft size. It is expected that an access shaft will range from 8 to 20 ft in diameter.

### Non-mining Methods

Whereas conventional crude oils may have a viscosity of several poise (at 40°C, 105°F), the tar sand bitumen has a viscosity of the order of 50,000–1,000,000 cP or more at formation temperatures (approximately 0–10°C, 32–50°F depending upon the season). This offers a formidable (but not insurmountable) obstacle to bitumen recovery.

In principle, the *non-mining recovery of bitumen from tar sand deposits* is an enhanced recovery technique and requires the injection of a fluid into the formation through an injection wall. This leads to the in situ displacement of the bitumen from the sand

followed by bitumen production at the surface through an egress well (production well).

In tar sand deposits, it is often desirable to initiate *enhanced oil recovery* (EOR) operations as early as possible, which mean considerably abbreviating conventional secondary recovery operations or bypassing them altogether. Thermal floods using steam and controlled in situ combustion methods are also used. Thermal methods of recovery reduce the viscosity of the crude oil by heat so that it flows more easily into the production well [28].

The technologies applied to oil recovery involve different concepts, some of which can cause changes to the oil during production. Technologies such as alkaline flooding, microemulsion (micellar/emulsion) flooding, polymer-augmented waterflooding, and carbon dioxide miscible/immiscible flooding do not require or cause any change to the oil. The steaming technologies may cause some steam distillation that can augment the process when the steam-distilled material moves with the steam front and acts as a solvent for oil ahead of the steam front. Again, there is no change to the oil although there may be favorable compositional changes to the oil insofar as lighter fractions are recovered and heavier materials remain in the reservoir.

The technology where changes do occur involves combustion of the oil in situ. The concept of any combustion technology requires that the oil be partially combusted and that thermal decomposition occur to other parts of the oil. This is sufficient to cause irreversible chemical and physical changes to the oil to the extent that the product is markedly different to the oil-in-place, indicating upgrading of the bitumen during the process. Recognition of this phenomenon is essential before combustion technologies are applied to oil recovery.

*Thermal recovery methods* (Fig. 1) have found most use when heavy oil or bitumen has an extremely high viscosity under reservoir conditions [2, 4]. For example, bitumen is highly viscous, with a viscosity ranging up to a million centipoises or more at the reservoir conditions.

*Thermal-enhanced oil recovery processes* (i.e., cyclic steam injection, steam flooding, and in situ combustion) add heat to the reservoir to reduce oil viscosity and/or to vaporize the oil. In both instances, the oil is

made more mobile so that it can be more effectively driven to producing wells. In addition to adding heat, these processes provide a driving force (pressure) to move oil to producing wells.

In the *modified in situ extraction* processes, combinations of in situ and mining techniques are used to access the reservoir. A portion of the reservoir rock must be removed to enable application of the in situ extraction technology. The most common method is to enter the reservoir through a large-diameter vertical shaft, excavate horizontal drifts from the bottom of the shaft, and drill injection and production wells horizontally from the drifts. Thermal extraction processes are then applied through the wells. When the horizontal wells are drilled at or near the base of the tar sand reservoir, the injected heat rises from the injection wells through the reservoir, and drainage of produced fluids to the production wells is assisted by gravity.

There are, however, several serious constraints that are particularly important and relate to bulk properties of the tar sand and the bitumen. In fact, both must be considered in the context of bitumen recovery by non-mining techniques. For example, the Canadian deposits are unconsolidated sands with a porosity ranging up to about 45% whereas other deposits may range from predominantly low-porosity, low-permeability consolidated sand to, in a few instances, unconsolidated sands. In addition, the bitumen properties are not conducive to fluid flow under deposit conditions. Nevertheless, where the general nature of the deposits prohibits the application of a mining technique, a non-mining method may be the only feasible bitumen recovery option.

Another general constraint to bitumen recovery by non-mining methods is the relatively low injectivity of tar sand formations. Thus, it is usually necessary to inject displacement or recovery fluids at a pressure such that fracturing (parting) is achieved. Such a technique therefore changes the reservoir profile and introduces a series of channels through which fluids can flow from the injection well to the production well. On the other hand, the technique may be disadvantageous insofar as the fracture occurs along the path of least resistance, giving undesirable (i.e., inefficient) flow characteristics within the reservoir between the injection and production wells, leaving

a large part of the reservoir relatively untouched by the displacement or recovery fluids.

Another general constraint to bitumen recovery by non-mining methods is the relatively low injectivity of tar sand formations. It is usually necessary to inject displacement/recovery fluids at a pressure such that fracturing (parting) is achieved. Such a technique, therefore, changes the reservoir profile and introduces a series of channels through which fluids can flow from the injection well to the production well. On the other hand, the technique may be disadvantageous insofar as the fracture occurs along the path of least resistance giving undesirable (i.e., inefficient) flow characteristics within the reservoir between the injection and production wells which leave a part of the reservoir relatively untouched by the displacement or recovery fluids.

**Steam-Based Processes** Steam-based processes are the most advanced of all enhanced oil recovery methods in terms of field experience and thus have the least uncertainty in estimating performance, provided that a good reservoir description is available. Steam processes are most often applied in reservoirs containing heavy oil, which is mobile at reservoir temperature. Commercial application of steam processes has been underway since the early 1960s.

*Steam drive injection (steam injection)* has been commercially applied since the early 1960s. The process occurs in two steps: (1) steam stimulation of production wells, that is, direct steam stimulation, and (2) steam drive by steam injection to increase production from other wells (i.e., indirect steam stimulation). Steam drive requires sufficient effective permeability (with the immobile bitumen in place) to allow injection of the steam at rates sufficient to raise the reservoir temperature to mobilize the bitumen and drive it to the production well.

*Cyclic steam injection* (also called *huff and puff* or *steam soak*) is the alternating injection of steam and production of oil with condensed steam from the same well or wells. This process is predominantly a vertical well process, with each well alternately injecting steam and producing heavy oil and steam condensate. In practice, steam is injected into the formation at greater than fracturing pressure followed by a *soak* period after which production is commenced. The heat injected warms the heavy oil and lowers its viscosity. A heated

zone is created through which the warmed heavy oil can flow back into the well. This is a well-developed process; the major limitation is that less than 30% (usually less than 20%) of the initial oil-in-place can be recovered.

**Combustion Processes** In situ *combustion (fireflood)* is normally applied to reservoirs containing low-gravity oil but has been tested over perhaps the widest spectrum of conditions of any enhanced oil recovery process. In the process, heat is generated within the reservoir by injecting air and burning part of the crude oil. This reduces the oil viscosity and partially vaporizes the oil-in-place, and the oil is driven out of the reservoir by a combination of steam, hot water, and gas drive. Forward combustion involves movement of the hot front in the same direction as the injected air; reverse combustion involves movement of the hot front opposite to the direction of the injected air.

During the process, energy is generated in the formation by igniting bitumen in the formation and sustaining it in a state of combustion or partial combustion. The high temperatures generated decrease the viscosity of the oil and make it more mobile. Some cracking of the bitumen also occurs, and an upgraded product rather than bitumen itself is the fluid recovered from the production wells.

The relatively small portion of the oil that remains after the displacement mechanisms have acted becomes the fuel for the in situ combustion process. Production is obtained from wells offsetting the injection locations. In some applications, the efficiency of the total in situ combustion operation can be improved by alternating water and air injection. The injected water tends to improve the utilization of heat by transferring heat from the rock behind the combustion zone to the rock immediately ahead of the combustion zone.

The use of combustion to stimulate oil production is regarded as attractive for deep reservoirs. In contrast to steam injection, it usually involves no loss of heat. The duration of the combustion may be less than 30 days or much as 90 days depending on requirements. In addition, backflow of the oil through the hot zone must be prevented or coking will occur.

*Forward combustion* involves movement of the hot front in the same direction as the injected air while reverse combustion involves movement of the hot front

opposite to the direction of the injected air. In forward combustion, the hydrocarbon products released from the zone of combustion move into a relatively cold portion of the formation. Thus, there is a definite upper limit of the viscosity of the liquids that can be recovered by a forward-combustion process. On the other hand, since the air passes through the hot formation before reaching the combustion zone, burning is complete; the formation is left completely cleaned of hydrocarbons.

*Reverse combustion* is particularly applicable to reservoirs with lower effective permeability (in contrast with forward combustion). It is more effective because the lower permeability would cause the reservoir to be plugged by the mobilized fluids ahead of a forward-combustion front. In the reverse combustion process, the vaporized and mobilized fluids move through the heated portion of the reservoir behind the combustion front. The reverse combustion partially cracks the bitumen, consumes a portion of the bitumen as fuel, and deposits residual coke on the sand grains. In the process, part of the bitumen will be consumed as fuel and part will be deposited on the sand grains as coke leaving 40–60% recoverable. This coke deposition serves as a cementing material, reducing movement and production of sand.

The addition of water or steam to an in situ combustion process can result in a significant increase in the overall efficiency of that process. Two major benefits may be derived. Heat transfer in the reservoir is improved because the steam and condensate have greater heat-carrying capacity than combustion gases and gaseous hydrocarbons. Sweep efficiency may also be improved because of the more favorable mobility ratio of steam-bitumen compared with gas-bitumen.

Applying a preheating phase before the bitumen recovery phase may significantly enhance the steam or combustion extraction processes. Preheating can be particularly beneficial if the saturation of highly viscous bitumen is sufficiently great as to lower the effective permeability to the point of production being precluded by reservoir plugging. Preheating partially mobilizes the bitumen by raising its temperature and lowering its viscosity. The result is a lower required pressure to inject steam or air and move the bitumen.

In the *fracture-assisted steam technology* (FAST) process, steam is injected rapidly into an induced

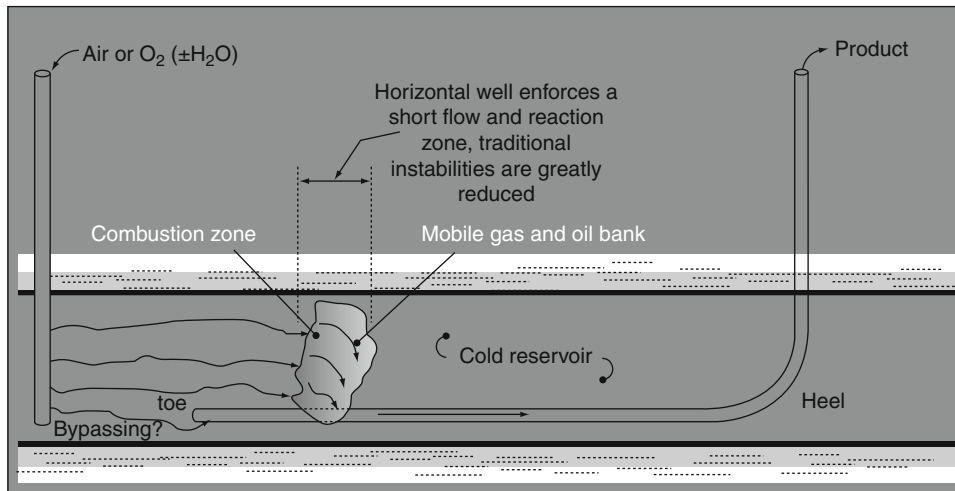
horizontal fracture near the bottom of the reservoir to preheat the reservoir. This process has been applied successfully in three pilot projects in southwest Texas. Shell has accomplished the same preheating goal by injecting steam into a high-permeability bottom water zone in the Peace River (Alberta) field. Electrical heating of the reservoir by radio-frequency waves may also be an effective method.

In situ combustion has been field tested under a wide variety of reservoir conditions, but few projects have proven economical and advanced to commercial scale and the concept has been abandoned by many recovery operators. However, in situ combustion may make a comeback with a new concept. THAI (toe-to-heel air injection) (Fig. 3) is based on the geometry of horizontal wells that may solve the problems that have plagued conventional in situ combustion. The well geometry enforces a short flow path so that any instability issues associated with conventional combustion are reduced or even eliminated [2, 4].

In situ conversion, or underground refining, is a promising new technology to tap the extensive reservoirs of heavy oil and deposits of bitumen. The new technology [29, 30] features the injection of high-temperature, high-quality steam, and hot hydrogen into a formation containing heavy hydrocarbons to initiate conversion of the heavy hydrocarbons into lighter hydrocarbons. In effect, the heavy hydrocarbons undergo partial underground refining that converts them into a synthetic crude oil (or *syncrude*). The heavier portion of the syncrude is treated to provide the fuel and hydrogen required by the process, and the lighter portion is marketed as a conventional crude oil.

Thus, below ground, superheated steam and hot hydrogen are injected into a heavy oil or bitumen formation, which simultaneously produces the heavy oil or bitumen and converts it in situ (i.e., within the formation) into syncrude. Above ground, the heavier fraction of the syncrude is separated and treated on-site to produce the fuel and hydrogen required by the process, while the lighter fraction is sent to a conventional refinery to be made into petroleum products (United States Patent 6,016,867; United States Patent 6,016,868).

The potential advantages of an in situ process for bitumen and heavy oil include (1) leaving the



**Petroleum and Oil Sands Exploration and Production. Figure 3**  
The THAI process

carbon-forming precursors in the ground, (2) leaving the heavy metals in the ground, (3) reducing sand handling, and (4) bringing a partially upgraded product to the surface. The extent of the upgrading can, hopefully, be adjusted by adjusting the exposure of the bitumen of heavy oil to the underground thermal effects.

Finally, by all definitions, the quality of the bitumen from tar sand deposits is poor when considered as a refinery feedstock. As in any field in which primary recovery operations are followed by secondary or enhanced recovery operations and there is a change in product quality, such is also the case for tar sand recovery operations. Thus, product oils recovered by the thermal stimulation of tar sand deposits show some improvement in properties over those of the bitumen in-place.

In situ recovery processes (although less efficient in terms of bitumen recovery relative to mining operations) may have the added benefit of *leaving* some of the more obnoxious constituents (from the processing objective) in the ground.

**Other Processes** Many innovative concepts in heavy oil production have been developed in the last 10 years [2, 4]. However, there are varying degrees of success and all are dependent on the properties of the deposit. There is no panacea for bitumen recovery that can be applied on a worldwide basis.

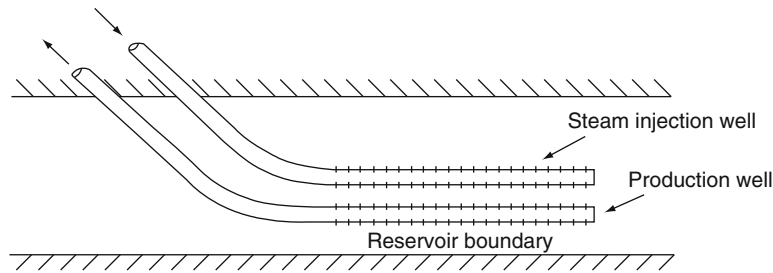
Long *horizontal wells* with several multilateral branches have been used widely in the development

of the heavy oils of Venezuela, where production rates as high as 2,000–2,500 bbl/day in some wells have been achieved through the use of aggregate horizontal lengths as large as 10,000 m in oil of 1,200–5,000 cP viscosity. Unfortunately, this technology can only achieve 8–15% v/v recovery of the oil and only from the best high-permeability zones.

*Inert gas injection (IGI)* is a technology for conventional oils in reservoirs where good vertical permeability exists, or where it can be created through propped hydraulic fracturing. It is generally viewed as a *top-down* process with nitrogen or methane injection through vertical wells at the top of the reservoirs, creating a gas–oil interface that is slowly displaced toward long horizontal production wells [2, 4]. As with all gravity drainage processes, it is essential to balance the injection and production volumes precisely so that the system does not become pressure driven, but remains in the gravity-dominated flow regime.

SAGD (steam-assisted gravity drainage) (Fig. 4) was developed first in Canada for reservoirs where the immobile bitumen occurs [31]. This process uses paired horizontal wells. Low-pressure steam continuously injected through the upper well creates a steam chamber along the walls of which the heated bitumen flows and is produced in the lower well.

In the process, a pair of horizontal wells that are separated vertically by about 15–20 ft are drilled at the bottom of a thick unconsolidated sandstone reservoir.



**Petroleum and Oil Sands Exploration and Production. Figure 4**

The SAGD process

Steam, perhaps along with a mixture of hydrocarbons that dissolve into the oil and help reduce its viscosity, is injected into the upper well. The heat reduces the oil viscosity to values as low as 1–10 cP (depending on temperature and initial conditions) and develops a *steam chamber* that grows vertically and laterally. The steam and gases rise because of their low density, and the oil and condensed water are removed through the lower well. The gases produced during SAGD tend to be methane with some carbon dioxide and traces of hydrogen sulfide.

The SAGD process, as for all gravity-driven processes, is extremely stable because the process zone grows only by gravity segregation, and there are no pressure-driven instabilities such as channeling, coning, and fracturing. SAGD seems to be relatively insensitive to shale streaks and similar horizontal barriers, even up to several meters thick (3–6 ft), that otherwise would restrict vertical flow rates. The combined processes of gravity segregation and shale thermal fracturing make SAGD so efficient that recovery ratios of 60–70% are claimed. Nevertheless, the process is not universally applicable to all reservoirs and deposits.

*Cold heavy oil production with sand* (CHOPS) is also used as a production approach in unconsolidated sandstones. The process results in the development of high-permeability channels (*wormholes*) in the adjacent low cohesive strength sands, facilitating the flow of oil foam that is caused by solution gas drive. Instead of blocking sand ingress by screens or gravel packs, sand is encouraged to enter the wellbore by aggressive perforation and swabbing strategies. Vertical or slightly inclined wells (vertical to 45°) are operated with rotary progressive cavity pumps (rather than reciprocating pumps) and

old fields are converting to higher-capacity progressive cavity pumps, giving production boosts to old wells. Because massive sand production creates a large disturbed zone, the reservoir may be positively affected for later implementation of thermal processes.

Typically, a well placed on CHOPS production will initially produce a high percentage of sand, greater than 20% by volume of liquids. However, this generally drops after some weeks or months. The huge volumes of sand are disposed of by slurry fracture injection or salt cavern placement or by sand placement in a landfill in an environmentally acceptable manner. Obviously, the production of excessive amounts of sand is a cause for mechanical and environmental concern.

*Pressure pulsing technologies* (PPT) involves a radically new aspect of porous media mechanics discovered and developed into a production enhancement method in the period 1997–2003. The mechanism by which PPT works is to generate a porosity dilation wave (a fluid displacement wave similar to a tsunami); this generates pore-scale dilation and contraction so that oil and water flow into and out of pores, leading to periodic fluid accelerations in the pore throats. As the porosity dilation wave moves through the porous medium at a velocity of about 50–100 ft/s (40–80 m/s), the small expansion and contraction of the pores with the passage of each packet of wave energy helps unblock pore throats, increase the velocity of liquid flow, overcome part of the effects of capillary blockage, and reduce some of the negative effects of instability due to viscous fingering, coning, and permeability streak channeling.

PPT promises to be a major adjunct to a number of oil production processes, particularly all pressure-driven processes, where it will both accelerate flow



rates as well as increasing oil recovery factors. It is also now used in environmental applications to help purge shallow aquifers of nonmiscible phases such as oil. The basis for its use in tar sand deposits is largely unknown and unproven.

*Vapor-assisted petroleum extraction* (VAPEX) is a new process in which the physics of the process are essentially the same as for SAGD and the configuration of wells is generally similar. The process involves the injection of vaporized solvents such as ethane or propane to create a vapor-chamber through which the oil flows due to gravity drainage [32–35]. The process can be applied in paired horizontal wells, single horizontal wells, or a combination of vertical and horizontal wells. The key benefits are significantly lower energy costs, potential for in situ upgrading, and application to thin reservoirs, with bottom water or reactive mineralogy.

Because of the slow diffusion of gases and liquids into viscous oils, this approach, used alone, perhaps will be suited only for less viscous oils although preliminary tests indicate that there are micro-mechanisms that act so that the VAPEX dilution process is not diffusion rate limited and the process may be suitable for the highly viscous tar sand bitumen [36, 37].

Nevertheless, VAPEX can undoubtedly be used in conjunction with SAGD methods. As with SAGD and IGI, a key factor is the generation of a three-phase system with a continuous gas phase so that as much of the oil as possible can be contacted by the gaseous phases, generating the thin oil film drainage mechanism. As with IGI, vertical permeability barriers are a problem, and must be overcome through hydraulic fracturing to create vertical permeable channels, or undercut by the lateral growth of the chamber beyond the lateral extent of the limited barrier, or “baffle.” As with any solvent process, the loss of solvents in geological formations (such as by adsorption on clay and other minerals) drastically affects process economics and raises many serious environmental issues.

Hybrid approaches that involve the simultaneous use of several technologies are evolving and will see greater applications in the future. In addition to hybrid approaches, the new production technologies, along with older, pressure-driven technologies, will be used in successive phases to extract more oil from reservoirs,

even from reservoirs that have been abandoned after primary exploitation. These hybrid approaches hold (on paper at least) the promise of significantly increasing recoverable reserves worldwide, not just in heavy oil cases.

*Microbial-enhanced oil recovery* (MEOR) processes involve use of reservoir microorganisms or specially selected natural bacterial to produce specific metabolic events that lead to enhanced oil recovery.

In microbial-enhanced oil recovery processes, microbial technology is exploited in oil reservoirs to improve recovery [38–40]. From a microbiologist’s perspective, microbial-enhanced oil recovery processes are somewhat akin to in situ bioremediation processes. Injected nutrients, together with indigenous or added microbes, promote in situ microbial growth and/or generation of products which mobilize additional oil and move it to producing wells through reservoir repressurization, interfacial tension/oil viscosity reduction, and selective plugging of the most permeable zones [41, 42].

This technology requires consideration of the physicochemical properties of the reservoir in terms of salinity, pH, temperature, pressure, and nutrient availability [43, 44].

The microbial-enhanced oil recovery process may modify the immediate reservoir environment in a number of ways that could also damage the production hardware or the formation itself. Certain sulfate reducers can produce  $H_2S$ , which can corrode pipeline and other components of the recovery equipment, and considerable uncertainty still remains regarding process performance. In addition, conditions vary from reservoir to reservoir, which calls for reservoir-specific customization of the microbial-enhanced oil recovery process, and this alone has the potential to undermine microbial process economic viability. Even though microbes produce the necessary chemical reactions in situ, there is need for caution and astute observation of the effects of the microorganisms on the reservoir chemistry.

Finally, recent developments in *upgrading* of heavy oil and bitumen [2, 4] indicate that the near future could see a reduction of the differential cost of upgrading heavy oil. These processes are based on a better understanding of the issues of asphaltene solubility effects at high temperatures, incorporation of

a catalyst that is chemically precipitated internally during the upgrading, and improving hydrogen addition or carbon rejection.

### Future Directions

With the current energy problems, the motivation for recovering as much as possible of the in-place reserves is greater than ever. There is a potentially uncomfortable and politically disastrous situation where the gap between energy requirements and available energy supplies is widening quickly.

Consequently, the search for new domestic supplies has shifted in large measure to increasingly hostile environments such as the Alaskan North Slope and the offshore waters along the outer continental shelves of the United States. These changes in production operations (not to mention the associated environmental disasters that often accompany such venture) have meant both significantly higher costs of production operations and fewer and fewer new commercial discoveries.

The importance of improving the rate of recovery from domestic petroleum reservoirs is underscored by the increasing difficulty of finding significant new reserves to meet the increasing demand for energy. One solution lies in greater emphasis on a multidisciplinary approach – on an intracompany basis and on a cooperative intercompany basis within the industry.

In the near term, a more immediate solution lies in improved application of existing technology as regards selection and quality control of materials, rigorous application of procedures, and the training and supervision of personnel. Part of the answer to the shortage, at least for the short term, has been to import more crude oil, but this is a less than ideal solution for a number of economic and political reasons.

In terms of petroleum recovery, steam-based processes will remain the processes of choice for the recovery of much of the oil in the ground over the next 2 or 3 decades. In some instances, fire flooding will be rejuvenated as the need to recovery of bitumen and residual oil becomes more important.

With the preponderance of heavier oils and tars and bitumen, it is likely that efforts will be made to emphasize partial (or full) upgrading in situ as an integral part

of the recovery process. Any type of upgrading during recovery will enhance the quality of the recovered oil, leaving some of the undesirable constituents in the ground as thermal products. Enhancement of the quality of the recovered oil will facilitate upgrading of the oil in the refinery.

Biotechnology will play a more significant role in enhancing crude oil recovery from the depleted oil reservoirs to solve stagnant petroleum production. Such enhanced oil recovery processes (microbial-enhanced oil recovery, MEOR) involve stimulating indigenous reservoir microbes or injecting specially selected consortia of natural bacteria into the reservoir to produce specific metabolic events that lead to improved oil recovery. This also involves flooding with oil recovery agents produced *ex situ* by industrial or pilot scale fermentation. However, like all recovery processes, a complete evaluation and assessment of microbial from a scientific and engineering standpoint must be performed and must be based on economics, applicability, and the performance standards required to further improve the process efficiency.

Above all, there is the need to manage recovery operations in such a manner that environmental issues do not become issues!

### Bibliography

#### Primary Literature

1. Taber JJ, Martin FD (1983) Technical screening guides for the enhanced recovery of oil. In: Proceedings of the 58th SPE annual technical conference and exhibition, San Francisco, 5–8 Oct 1983 (SPE 12069)
2. Speight JG (2009) Enhanced recovery methods for heavy oil and tar sands. Gulf Publishing, Houston
3. Drew LJ (1997) Undiscovered mineral and petroleum deposits: assessment & controversy. Plenum, New York (Chap. 3)
4. Speight JG (2007) The chemistry and technology of petroleum, 4th edn. CRC Press/Taylor and Francis, Boca Raton
5. Lake LW, Walsh MP (2004) Primary hydrocarbon recovery. Elsevier, Amsterdam
6. Craft BC, Hawkins MF (1959) Applied petroleum reservoir engineering. Prentice-Hall, Englewood Cliffs
7. Frick TC (1962) Petroleum production handbook, vol II. McGraw-Hill, New York
8. Lake LW (1989) Enhanced oil recovery. Prentice-Hall, Englewood Cliffs
9. Arnarnath A (1999) Enhanced oil recovery scoping study. Report No. TR-113836, Electric Power Research Institute, Palo Alto

10. Borchardt JK, Yen TF (1989) Oil field chemistry, Symposium series No. 396. American Chemical Society, Washington, DC
11. CFR 1.43-2 (2004) Internal revenue service, Department of the Treasury, Government of the United States, Washington, DC
12. OTA (1978) Enhanced oil recovery potential in the United States. Office of Technology Assessment, Washington, DC. NTIS order #PB-276594
13. Reed RL, Healy RN (1977) Some physicochemical aspects of microemulsion flooding: a review. In: Shah DO, Schechter RS (eds) Improved oil recovery by surfactant and polymer flooding. Academic, New York
14. Speight JG (2000) Desulfurization of heavy oils and residua, 2nd edn. Marcel Dekker, New York
15. Terwilliger PL (1975) Paper 5568. In: Proceedings of 50th annual fall meeting of the Society of Petroleum Engineers, Dallas. American Institute of Mechanical Engineers, Dallas
16. Terwilliger PL, Clay RR, Wilson LA, Gonzalez-Gerth E (1975) J Petrol Technol 27:9
17. Burger J (1978) Developments in petroleum science. In: Chilingarian GV, Yen TF (eds) Bitumens, asphalts and tar sands, vol 7. Elsevier, New York, p 191
18. Clark KA (1944) Hot-water separation of Alberta bituminous sand. Trans Can Inst Min Met 47:257
19. Carrigy MA (1963) Bulletin No. 14. Alberta Research Council, Edmonton
20. Carrigy MA (1963) The oil sands of Alberta. Information Series No. 45. Alberta Research Council, Edmonton
21. Fear JVD, Innes, ED (1967) In: Proceedings seventh world petroleum congress, Mexico, vol 3, p 549
22. Speight JG, Moschopedis SE (1978) Fuel Process Technol 1:261
23. Miller JC, Misra M (1982) Fuel Process Technol 6:27
24. Misra M, Aguilar R, Miller JD (1981) Sep Sci Technol 16(10):1523
25. Kruyer J (1982) In: Proceedings of the second international conference on heavy crude and tar sands, Caracas, 7–17 Feb 1982
26. Kruyer J (1983) Preprint No. 3d. Summer national meeting of the American Institute of Chemical Engineers, Denver, 28–31 Aug 1983
27. Gishler PE (1949) Can J Res 27:104
28. Pratts M (1986) Thermal recovery. Society of Petroleum Engineers, New York
29. Gregoli AA, Rimmer DP, Graue DJ (2000) Upgrading and recovery of heavy crude oils and natural bitumen by in situ hydrovisbreaking. US Patent 6,016,867, 25 Jan 2000
30. Gregoli AA, Rimmer DP (2000) Production of synthetic crude oil from heavy hydrocarbons recovered by in situ hydrovisbreaking. US Patent 6,016,868, 25 Jan 2000
31. Dusseault MB, Geilikman MB, Spanos TJT (1998) J Petrol Technol 50(9):92–94
32. Butler RM, Mokrys IJ (1991) J Can Pet Technol 30(1):97–106
33. Butler RM, Mokrys IJ (1995) Process and apparatus for the recovery of hydrocarbons from a hydrocarbon deposit. US Patent 5,407,009, 18 Apr 1995
34. Butler RM, Mokrys IJ (1995) Process and apparatus for the recovery of hydrocarbons from a hydrocarbon deposit. US Patent 5,607,016, 4 Mar 1995
35. Butler RM, Jiang Q (2000) J Can Pet Technol 39:48–56
36. Yang C, Gu Y (2005a) A novel experimental technique for studying solvent mass transfer and oil swelling effect in a vapor extraction (VAPEX) process. Paper No. 2005–099. In: Proceedings of the 56th annual technical meeting. The Canadian international petroleum conference, Calgary, 7–9 Jun 2005
37. Yang C, Gu Y (2005b) Effects of solvent-heavy oil interfacial tension on gravity drainage in the VAPEX process. Paper No. SPE 97906. In: Society of petroleum engineers international thermal operations and heavy oil symposium, Calgary, 1–3 Nov 2005
38. Banat IM (1995) Biosurfactant production and possible uses in microbial enhanced oil recovery and oil pollution remediation. Bioresour Technol 51:1–12
39. Clark JB, Munnecke DM, Jenneman GE (1981) In situ microbial enhancement of oil production. Dev Ind Microbiol 15:695–701
40. Stosur GJ (1991) Unconventional EOR concepts. Crit Rep Appl Chem 33:341–373
41. Bryant RS, Lindsey RP (1996) World-wide applications of microbial technology for improving oil recovery. In: Proceedings of the SPE symposium on improved oil recovery of the society of petroleum engineers, Richardson, pp 27–134
42. Bryant RS, Donaldson EC, Yen TF, Chilingarian GV (1989) Microbial enhanced oil recovery. In: Donaldson EC, Chilingarian GV, Yen TF (eds) Enhanced oil recovery II: processes and operations. Elsevier, Amsterdam, pp 423–450
43. Khire JM, Khan MI (1994) Microbially enhanced oil recovery (MEOR). Part 1. Importance and mechanisms of microbial enhanced oil recovery. Enzyme Microb Technol 16: 170–172
44. Khire JM, Khan MI (1994) Microbially enhanced oil recovery (MEOR). Part 2. Microbes and the subsurface environment for microbial enhanced oil recovery. Enzyme Microb Technol 16:258–259

## Books and Reviews

- Ancheyta J, Speight JG (2007) Hydroprocessing of heavy oils and residua. CRC Press/Taylor & Francis, Boca Raton
- Bower T (2009) Oil: money, politics, and power in the 21st century. Grand Central Publishing, Hachette Book Group, New York
- Gudmestad OT, Zolotukhin AB, Jarlsby ET (2010) Petroleum resources, with emphasis on offshore fields. WIT Press, Billerica
- Nersesian RL (2010) Energy for the 21st century: a comprehensive guide to conventional and alternative sources, 2nd edn. M.E. Sharpe, Armonk
- Speight JG (2008) Synthetic fuels handbook: properties, processes, and performance. McGraw-Hill, New York
- Wihbet PM (2009) The rise of the new oil order. Academy & Finance, Geneva

## Petroleum Refining and Environmental Control and Environmental Effects

JAMES G. SPEIGHT

CD&W Inc, Laramie, WY, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Definitions

Environmental Regulations

Processes and Process Wastes

Future Directions

Bibliography

### Glossary

**Emissions** Gaseous, liquid, or solid by-products introduced into the environment as a result of refining processes.

**Environmental control** The use of various technologies to control and even prevent refinery emissions from entering the environment.

**Environmental effects** The effects of refinery emissions on the flora and fauna in the various ecosystems.

**Refining** The processes by which petroleum is distilled and/or converted by application of physical and chemical processes to form a variety of products.

**Regulations** The laws by which environmental emissions are controlled.

### Definition of the Subject

The work summarizes the various process emissions that occur during petroleum refining. There are also general descriptions of the various pollution, health, and environmental problems especially specific to the petroleum industry and places in perspective the government regulations as well as industry efforts to adhere to these regulations. The objective is to indicate the types of emissions and the laws that regulate these emissions.

### Introduction

Petroleum as an energy source use is a necessary part of the modern world and will be a primary source of energy for the next several decades, hence the need for control over the amounts and types of emissions from the use of petroleum and its products. Furthermore, the capacity of the environment to absorb the effluents and other impacts of process technologies is not unlimited and the environment should be considered to be an extremely limited resource, and discharge of chemicals into it should be subject to severe constraints – as a result, it is necessary to understand the nature and magnitude of the problems involved [1].

Both the production and processing of crude oil involve the use of a variety of substances [2], some toxic, including lubricants in oil wells and catalysts and other chemicals in refining (Fig. 1). The amounts used tend to be relatively easy to control, and the spillage of crude oil is more detrimental to the environment.

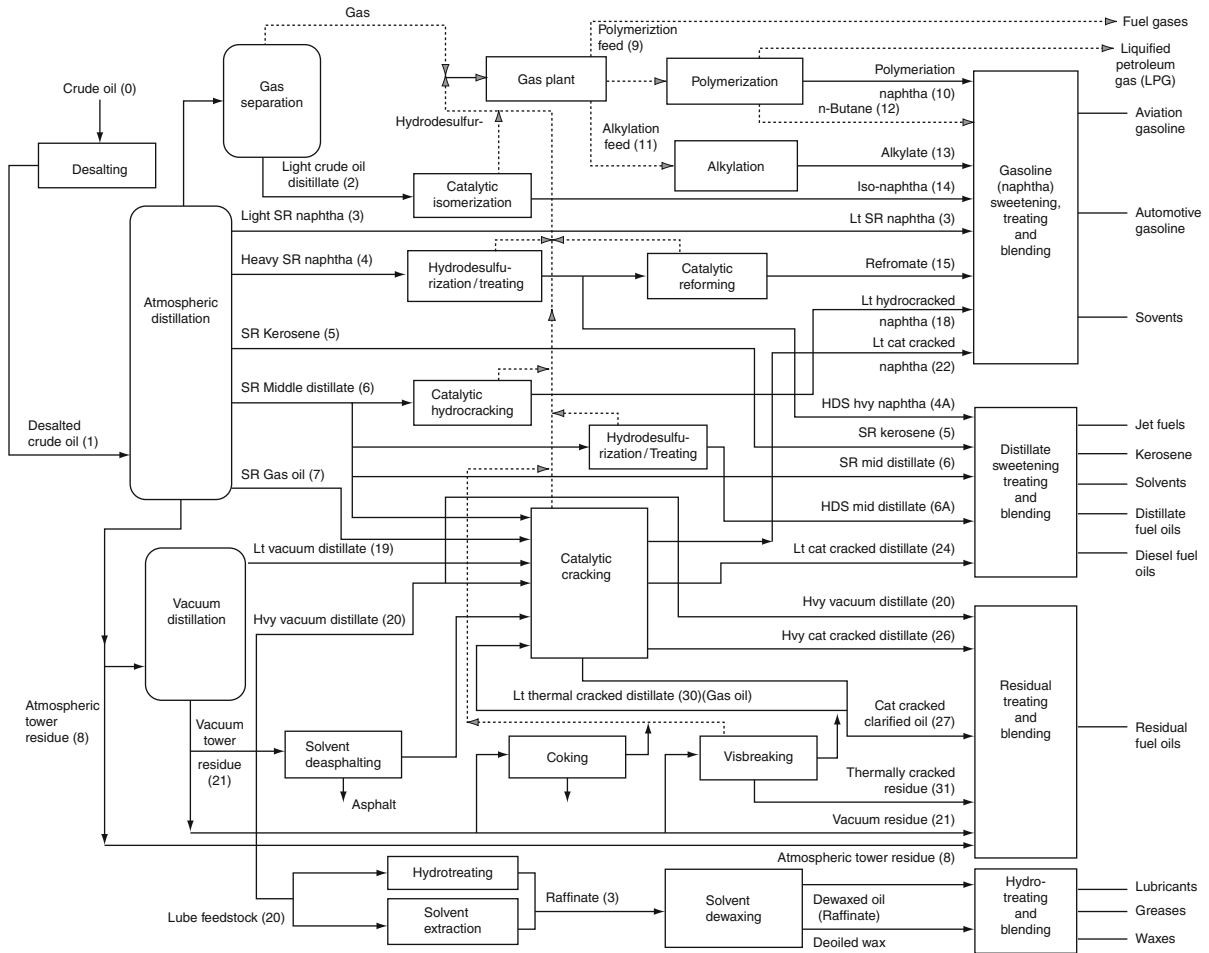
The purpose of this work is to summarize and generalize the various pollution, health, and environmental problems especially specific to the petroleum industry and to place in perspective government laws and regulations as well as industry efforts to control these problems [3–6]. The objective is to indicate the types of emissions and the laws that regulate these emissions.

### Definitions

Briefly, petroleum production and petroleum refining produce *chemical waste* [6]. If this *chemical waste* is not processed in a timely manner, it can become a *pollutant*. Under some circumstances, chemical waste is reclassified as *hazardous waste*.

*Hazardous waste* is any gaseous, liquid, or solid waste material that, if improperly managed or disposed of, may pose hazards to human health and the environment. In some cases, the term “*chemical waste*” is used interchangeably (often incorrectly) with the term “*hazardous waste*,” but chemical waste is always hazardous and the correct use of the terms must be used.

A *pollutant* is a substance present in a particular location (*ecosystem*) – usually it is not indigenous to the location or is present in a concentration greater than the concentration that occurs naturally. The substance



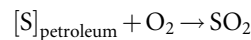
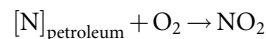
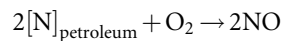
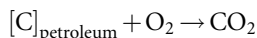
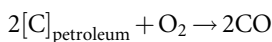
**Petroleum Refining and Environmental Control and Environmental Effects. Figure 1**

Schematic overview of a refinery (OSHA technical Manual, Section IV, Chapter 2: Petroleum Refining Processes, [http://www.osha.gov/dts/osta/otm/otm\\_iv/otm\\_iv\\_2.html](http://www.osha.gov/dts/osta/otm/otm_iv/otm_iv_2.html))

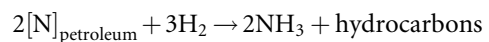
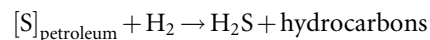
is often the product of human activity and has a detrimental effect on the environment, in part or in toto. Pollutants can also be subdivided into two classes: primary and secondary.

Source → Primary pollutant → Secondary pollutant

A *primary pollutant* is a pollutant that is emitted directly from the source. In terms of atmospheric pollutants from petroleum, examples are carbon oxides, sulfur dioxide, and nitrogen oxides from fuel combustion operations:

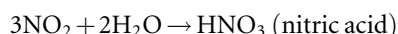
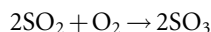


Hydrogen sulfide and ammonia are produced from processing sulfur-containing and nitrogen-containing feedstocks:



A *secondary pollutant* is a pollutant that is produced by the interaction of a primary pollutant with another chemical. A *secondary pollutant* may also be produced

by dissociation of a primary pollutant, or other effects within a particular ecosystem. Again, using the atmosphere as an example, the formation of the constituents of acid rain is an example of the formation of secondary pollutants:



In many cases, these secondary pollutants can have significant environmental effects, such as participation in the formation of acid rain and smog [5].

An *environmental regulation* is a legal mechanism that determines how the policy directives of an

environmental law are to be carried out. An *environmental policy* is a requirement that specifies operating procedures that must be followed. An *environmental guidance* is a document developed by a governmental agency that outlines a position on a topic or which gives instructions on how a procedure must be carried out. It explains how to do something and provides governmental interpretations on a governmental act or policy.

### Environmental Regulations

Environmental issues range from the effects of pollutants on the population at large to effects on the lives of workers in various occupations where sickness or disability can result from exposure to chemical agents [5, 7, 8].

There are a variety of regulations (Table 1) that apply to petroleum refining [6]. The most popular is

**Petroleum Refining and Environmental Control and Environmental Effects. Table 1** Environmental regulations that apply to energy production

	First enacted	Amended
Clean Air Act	1970	1977 1990
Clean Water Act (Water Pollution Control Act)	1948	1965 <sup>a</sup> 1972 <sup>b</sup> 1977 1987 <sup>c</sup>
Comprehensive Environmental Response, Compensation and Liability Act	1980	1986 <sup>d</sup>
Hazardous Material Transportation Act	1974	1990
Occupational Safety and Health Act	1970	1987 <sup>e</sup>
Oil Pollution Act	1924	1990 <sup>f</sup>
Resource Conservation and Recovery Act	1976	1980 <sup>g</sup>
Safe Drinking Water Act	1974	1986 <sup>h</sup>
Superfund Amendments and Reauthorization Act (SARA)	1986	
Toxic Substances Control Act	1976	1984 <sup>i</sup>

<sup>a</sup>Water Quality Act

<sup>b</sup>Water Pollution Control Act

<sup>c</sup>Water Quality Act

<sup>d</sup>SARA Amendments

<sup>e</sup>Several amendments during the 1980s

<sup>f</sup>Interactive with various water pollution acts

<sup>g</sup>Federal cancer policy initiated

<sup>h</sup>Several amendments during the 1970s and 1980s

<sup>i</sup>Import rule enacted

the series of regulations known as the Clean Air Act that first was introduced in 1967 and was subsequently amended in 1970 and most recently in 1990. The most recent amendments provide stricter regulations for the establishment and enforcement of national ambient air quality standards for, as an example, sulfur dioxide. These standards do not stand alone, and there are many national standards for sulfur emissions.

The laws of relevance to the petroleum industry are:

### **The Clean Air Act Amendments**

The first *Clean Air Act* of 1970 and the 1977 Amendments consisted of three titles. *Title I* dealt with stationary air emission sources, *Title II* with mobile air emission sources, and *Title III* with definitions of appropriate terms as well as applicable standards for judicial review.

The *Clean Air Act Amendments of 1990* contain extensive provisions for control of the accidental release of toxic substances from storage or transportation as well as the formation of acid rain (acid deposition). In addition, the requirement that the standards be technology based removes much of the emotional perception that all chemicals are hazardous as well as the guesswork from legal enforcement of the legislation. The requirement also dictates environmental and health protection with an ample margin of safety.

### **The Water Pollution Control Act (The Clean Water Act)**

There are several acts that relate to the protection of the waterways in the United States but of particular interest to the petroleum industry in the present context is the *Water Pollution Control Act (Clean Water Act)*. The objective of the Act is to restore and maintain the chemical, physical, and biological integrity of water systems.

The original Water Pollution Control Act of 1948 and The Water Quality Act of 1965 were generally limited to control of pollution of interstate waters and the adoption of water-quality standards by the states for interstate water within their borders. The first comprehensive water-quality legislation in the United States came into being in 1972 as the Water Pollution Control Act, which was amended in 1977 and retitled to become the Clean Water Act. Further

amendments in 1978 were enacted to deal more effectively with spills of crude oil, with other amendments following in 1987 under the new name of the Water Quality Act.

Section 311 of the Clean Water Act includes elaborate provisions for regulating intentional or accidental discharges of petroleum and of hazardous substances. Included are response actions required for oil spills and the release or discharge of toxic and hazardous substances. As an example, the person in charge of a vessel or an onshore or offshore facility from which any chemical substance is discharged, in quantities equal to or exceeding its reportable quantity, must notify the appropriate federal agency as soon as such knowledge is obtained. The *Exxon Valdez* disaster and the recent spillage of oil into the Gulf of Mexico by BP are well-known examples of such a discharge of chemicals – the chemical being oil.

### **The Safe Drinking Water Act**

The Safe Drinking Water Act, first enacted in 1974, was amended several times in the 1970s and 1980s to set national drinking water standards. The Act calls for regulations that (1) apply to public water systems, (2) specify contaminants that may have any adverse effect on the health of persons, and (3) specify contaminant levels. Statutory provisions are included to cover underground injection control systems. The Act also requires maximum levels at which a contaminant must have no known or anticipated adverse effects on human health, thereby providing an *adequate margin of safety*.

The Superfund Amendments and Reauthorization Act (SARA) set the same standards for groundwater as for drinking water in terms of necessary cleanup and remediation of an inactive site that might be a former petroleum refinery. Under the Act, all underground injection activities must comply with the drinking water standards as well as meet specific permit conditions that are in unison with the provisions of the Clean Water Act.

### **The Resource Conservation and Recovery Act**

Since its initial enactment in 1976, the Resource Conservation and Recovery Act (RCRA) continues to promote safer waste management programs. Besides the regulatory requirements for waste management, the

Act specifies the mandatory obligations of generators, transporters, and disposers of waste as well as those of owners and/or operators of waste treatment, storage, or disposal facilities. The waste might be garbage, refuse, and sludge from a treatment plant or from a water supply treatment plant or air pollution control facility and other discarded material, including solid, liquid, semisolid, or contained gaseous material resulting from industrial, commercial, mining, and agricultural operations and from community activities.

The Act also states that solid waste does not include solid, or dissolved, materials in domestic sewage, or solid or dissolved materials in irrigation return flows or industrial discharges. A solid waste becomes a hazardous waste if it exhibits any one of four specific characteristics: (1) ignitability, (2) reactivity, (3) corrosivity, or (4) toxicity. Certain types of solid wastes (e.g., household waste) are not considered to be hazardous, irrespective of their characteristics.

### **The Toxic Substances Control Act**

The Toxic Substances Control Act was first enacted in 1976 and was designed to provide controls for those chemicals that may threaten human health or the environment. Particularly hazardous are the cyclic nitrogen species and that often occur in high-boiling petroleum fractions, distillation residua, and cracked residua.

The Act specifies a *premanufacture notification* requirement by which any manufacturer must notify the Environmental Protection Agency at least 90 days prior to the production of a new chemical substance. Notification is also required even if there is a new use for the chemical that can increase the risk to the environment. No notification is required for chemicals that are manufactured in small quantities solely for scientific research and experimentation.

A *new chemical substance* is a chemical that is not listed in the Environmental Protection Agency Inventory of Chemical Substances or is an unlisted reaction product of two or more chemicals. In addition, the term “*chemical substance*” means any organic or inorganic substance of a particular molecular identity, including any combination of such substances occurring in whole or in part as a result of a chemical reaction or occurring in nature, and any element

or uncombined radical. The term “*mixture*” means any combination of two or more chemical substances if the combination does not occur in nature and is not, in whole or in part, the result of a chemical reaction.

### **The Comprehensive Environmental Response, Compensation, and Liability Act**

The Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA), generally known as *Superfund*, was first signed into law in 1980. The purpose of this Act is to provide a response mechanism for cleanup of any hazardous substance released, such as an accidental spill, or of a threatened release of a chemical.

Under this Act, a hazardous substance is any substance requiring (1) special consideration due to its toxic nature under the Clean Air Act, the Clean Water Act, or the Toxic Substances Control Act and (2) any waste that is hazardous waste under RCRA. Additionally, a pollutant or contaminant can be any other substance not necessarily designated by or listed in the Act but that *will or may reasonably* be anticipated to cause any adverse effect in organisms and/or their offspring.

### **The Occupational Safety and Health Act**

The *Occupational Safety and Health Administration* (OSHA) came into being in 1970 and is responsible for administering the Occupational Safety and Health Act. Occupational health hazards are those factors arising in or from the occupational environment that adversely impact health.

The goal of the Act is to ensure that employees do not suffer material impairment of health or functional capacity due to a lifetime occupational exposure to chemicals. The Act is also responsible for the means by which chemicals are contained.

### **The Oil Pollution Act**

The Oil Pollution Act of 1990 deals with pollution of waterways by crude oil. The Act specifically deals with petroleum vessels and onshore and offshore facilities and imposes strict liability for oil spills on their owners and operators.



## The Hazardous Materials Transportation Act

The Hazardous Materials Transportation Act authorizes the establishment and enforcement of hazardous material regulations for all modes of transportation by highway, water, and rail. The purpose of the Act is to ensure safe transportation of hazardous materials. The Act prevents any person from offering or accepting for transportation a hazardous material (any substance or material, including a hazardous substance and hazardous waste, which is capable of posing an unreasonable risk to health, safety, and property) for transportation anywhere within the United States.

The Act also imposes restrictions on the packaging, handling, and shipping of hazardous materials in which the appropriate documentation, markings, labels, and safety precautions are required.

## Processes and Process Wastes

Enhanced oil recovery (EOR) processes rely upon the use of chemical or thermal energy to recover crude oil that is trapped in pores of reservoir rock after primary and secondary (waterflood) crude oil production has ceased [2, 9].

Chemicals used for enhanced oil recovery include surfactants to reduce the interfacial tension between oil and water, and oil and rock interfaces. Many microorganisms produce biosurfactants and perform this activity by fermentation of inexpensive raw materials such as molasses. Several biosurfactants are being evaluated for use in enhanced oil recovery.

A major issue in enhanced oil recovery processes is the variation of permeability in petroleum reservoirs. When water is injected to displace oil, the water will preferentially flow through areas of highest permeability, and bypass much of the oil. When chemicals are injected, they may also flow preferentially into high-permeability zones with the water, but then will grow and block those zones. When high-permeability zones are blocked, sweep efficiency is improved, and thus oil recovery.

## Transportation

In addition to the conventional meaning of the term *process*, the *recovery* and *transportation* of petroleum also needs to be considered here.

Oil spills during petroleum *transportation* have been the most visible problem. There have also been instances of oil wells at sea “blowing out,” or flowing uncontrollably, although the amounts from blowouts tend to be smaller than from tanker accidents.

*Tanker accidents* typically have a severe impact on ecosystems because of the rapid release of hundreds of thousands of barrels of crude oil (or crude oil products) into a small area.

While oil is at least theoretically biodegradable, large-scale spills can overwhelm the ability of the ecosystem to break the oil down. Over time, the lighter portions of crude oil evaporate, leaving the nonvolatile portion. Oil itself breaks down the protective waxes and oils in the feathers and fur of birds and animals. Some crude oils contain toxic metals as well. The impact of any given oil spill is determined by the size of the spill, the degree of dispersal, and the chemistry of the oil. Spills at sea are thought to have a less detrimental effect than spills in shallow waters.

## Refining

Petroleum *refining* is a complex sequence of chemical events that result in the production of a variety of products (Fig. 1). In fact, petroleum refining might be considered as a collection of individual, yet related processes that are each capable of producing effluent streams [2, 6].

Petroleum refining, as it is currently known, will continue at least for the next 3 decades. In spite of the various political differences that have caused fluctuations in petroleum imports, it is reality that imports of petroleum and petroleum products into the United States are on the order of 67% of the total requirements [2].

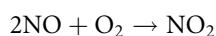
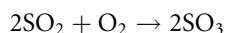
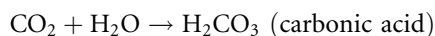
Petroleum, like any other raw material, is capable of producing chemical waste. By 1960, the petroleum refining industry had become well established throughout the world. Effluent water, atmospheric emissions, and combustion products also became a focus of increased technical attention [4, 5, 10–13].

Refineries produce a wide variety of products from petroleum feedstocks and feedstock blends [2, 14]. During petroleum refining, refineries use and generate an enormous amount of chemicals, some of which are present in air emissions, wastewater, or solid wastes (Table 2) [2, 6]. Emissions are also created through

**Petroleum Refining and Environmental Control and Environmental Effects. Table 2** Emissions and waste from refinery processes

Process	Air emissions	Residual wastes generated
Crude oil desalting	Heater stack gas (CO, SO <sub>x</sub> , NO <sub>x</sub> , hydrocarbons, and particulates), fugitive emissions (hydrocarbons)	Crude oil/desalter sludge (iron rust, clay, sand, water, emulsified oil and wax, metals)
Atmospheric distillation Vacuum distillation	Heater stack gas (CO, SO <sub>x</sub> , NO <sub>x</sub> , hydrocarbons and particulates), vents and fugitive emissions (hydrocarbons), steam ejector emissions (hydrocarbons), heater stack gas (CO, SO <sub>x</sub> , NO <sub>x</sub> , hydrocarbons, and particulates), vents and fugitive emissions (hydrocarbons)	Typically, little or no residual waste generated
Thermal cracking/ visbreaking	Heater stack gas (CO, SO <sub>x</sub> , NO <sub>x</sub> , hydrocarbons, and particulates), vents and fugitive emissions (hydrocarbons)	Typically, little or no residual waste generated
Coking	Heater stack gas (CO, SO <sub>x</sub> , NO <sub>x</sub> , hydrocarbons, and particulates), vents and fugitive emissions (hydrocarbons), and decoking emissions (hydrocarbons and particulates)	Coke dust (carbon particles and hydrocarbons)
Catalytic cracking	Heater stack gas (CO, SO <sub>x</sub> , NO <sub>x</sub> , hydrocarbons, and particulates), fugitive emissions (hydrocarbons), and catalyst regeneration (CO, NO <sub>x</sub> , SO <sub>x</sub> , and particulates)	Spent catalysts (metals from crude oil and hydrocarbons), spent catalyst fines from electrostatic precipitators (aluminum silicate and metals)
Catalytic hydrocracking	Heater stack gas (CO, SO <sub>x</sub> , NO <sub>x</sub> , hydrocarbons, and particulates), fugitive emissions (hydrocarbons), and catalyst regeneration (CO, NO <sub>x</sub> , SO <sub>x</sub> , and catalyst dust)	Spent catalysts fines
Hydrotreating/ hydroprocessing	Heater stack gas (CO, SO <sub>x</sub> , NO <sub>x</sub> , hydrocarbons, and particulates), vents and fugitive emissions (hydrocarbons), and catalyst regeneration (CO, NO <sub>x</sub> , SO <sub>x</sub> )	Spent catalyst fines (aluminum silicate and metals)
Alkylation	Heater stack gas (CO, SO <sub>x</sub> , NO <sub>x</sub> , hydrocarbons, and particulates), vents and fugitive emissions (hydrocarbons)	Neutralized alkylation sludge (sulfuric acid or calcium fluoride, hydrocarbons)
Isomerization	Heater stack gas (CO, SO <sub>x</sub> , NO <sub>x</sub> , hydrocarbons, and particulates), HCl (potentially in light ends), vents and fugitive emissions (hydrocarbons)	Calcium chloride sludge from neutralized HCl gas
Polymerization	H <sub>2</sub> S from caustic washing	Spent catalyst containing phosphoric acid
Catalytic reforming	Heater stack gas (CO, SO <sub>x</sub> , NO <sub>x</sub> , hydrocarbons, and particulates), fugitive emissions (hydrocarbons), and catalyst regeneration (CO, NO <sub>x</sub> , SO <sub>x</sub> )	Spent catalyst fines from electrostatic precipitators (alumina silicate and metals)
Solvent extraction	Fugitive solvents	Little or no residual wastes generated
Dewaxing	Fugitive solvents, heaters	Little or no residual wastes generated
Propane deasphalting	Heater stack gas (CO, SO <sub>x</sub> , NO <sub>x</sub> , hydrocarbons and particulates), fugitive propane	Little or no residual wastes generated
Wastewater treatment	Fugitive emissions (H <sub>2</sub> S, NH <sub>3</sub> , and hydrocarbons)	API separator sludge (phenols, metals and oil), chemical precipitation sludge (chemical coagulants, oil), DAF floats, biological sludge (metals, oil, suspended solids), spent lime

the combustion of fuels, and as by-products of chemical reactions occurring when petroleum fractions are upgraded. A large source of air emissions is, generally, the process heaters and boilers that produce carbon monoxide, sulfur oxides, and nitrogen oxides, leading to pollution and the formation of acid rain.



Hence, there is the need for gas-cleaning operations on a refinery site so that such gases are cleaned from the gas stream prior to entry into the atmosphere.

Fugitive emissions of volatile hydrocarbons arise from leaks in valves, pumps, flanges, and other similar sources where crude and its fractions flow through the system. While individual leaks may be minor, the combination of fugitive emissions from various sources can be substantial. These emissions are controlled primarily through leak detection and repair programs and occasionally through the use of special leak-resistant equipment.

In terms of individual processes, the potential for waste generation and, hence, leakage of emissions is as follows.

**Desalting** Petroleum often contains water, inorganic salts, suspended solids, and water-soluble trace metals. As a first step in the refining process, to reduce corrosion, plugging, and fouling of equipment and to prevent poisoning the catalysts in processing units, these contaminants must be removed by desalting (dehydration) [2].

The two most typical methods of petroleum desalting are (1) chemical separation and (2) electrostatic separation. In chemical desalting, water and chemical surfactant (demulsifiers) are added to the petroleum, heated so that salts and other impurities dissolve into the water or attach to the water, and then

held in a tank where they settle out. Electrical desalting is the application of high-voltage electrostatic charges to concentrate suspended water globules in the bottom of the settling tank. Surfactants are added only when the crude has a large amount of suspended solids. A third and less-common process involves filtering heated petroleum using diatomaceous earth.

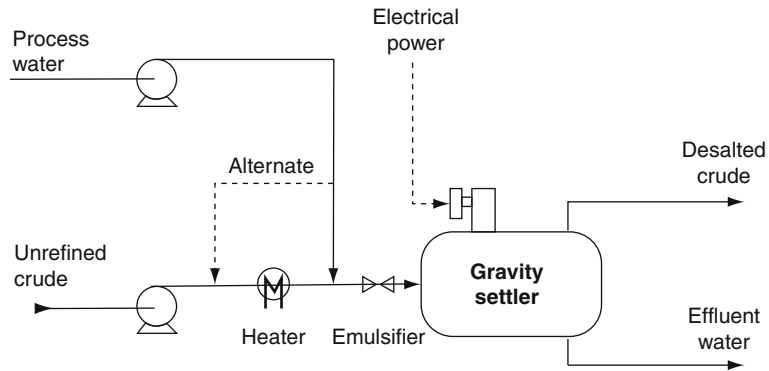
In the desalting process, the feedstock crude oil is heated to between 65°C and 177°C (150°F and 350°F) to reduce viscosity and surface tension for easier mixing and separation of the water, but the temperature is limited by the vapor pressure of the petroleum constituents. In both methods, other chemicals may be added – ammonia is often used to reduce corrosion and caustic or acid may be added to adjust the pH of the water wash. Wastewater and contaminants are discharged from the bottom of the settling tank to the wastewater treatment facility while the desalted crude is continuously drawn from the top of the settling tanks and sent to the crude distillation tower.

Desalting (Fig. 2) creates an oily desalter sludge that may be a hazardous waste and a high temperature salt wastewater stream (treated along with other refinery wastewaters). The primary polluting constituents in desalter wastewater include hydrogen sulfide, ammonia, phenol, high levels of suspended solids, and dissolved solids, with a high biochemical oxygen demand (BOD). In some cases, it is possible to recycle the desalter effluent water back into the desalting process, depending upon the type of crude being processed.

**Distillation** Atmospheric and vacuum distillation units (Figs. 3 and 4) are closed processes and exposures are expected to be minimal.

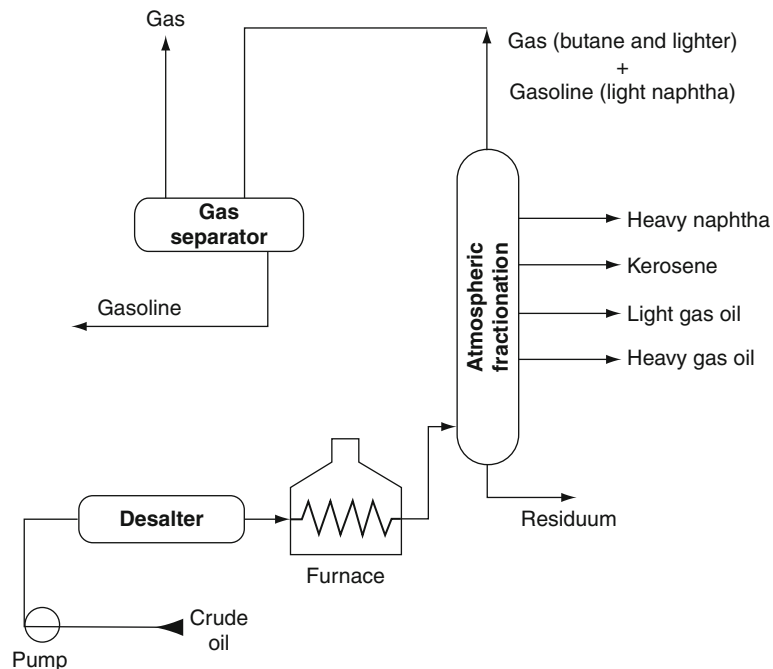
Both atmospheric distillation units and vacuum distillation units produce refinery fuel gas streams containing a mixture of light hydrocarbons, hydrogen sulfide, and ammonia. These streams are processed through gas treatment and sulfur recovery units to recover fuel gas and sulfur. Sulfur recovery creates emissions of ammonia, hydrogen sulfide, sulfur oxides, and nitrogen oxides.

When sour (high-sulfur) petroleum is processed, there is potential for exposure to hydrogen sulfide in the preheat exchanger and furnace, tower flash zone



Petroleum Refining and Environmental Control and Environmental Effects. Figure 2

Schematic of an electrostatic desalting unit (OSHA technical manual, Section IV, Chapter 2: Petroleum Refining Processes, [http://www.osha.gov/dts/osta/otm/otm\\_iv/otm\\_iv\\_2.html](http://www.osha.gov/dts/osta/otm/otm_iv/otm_iv_2.html))



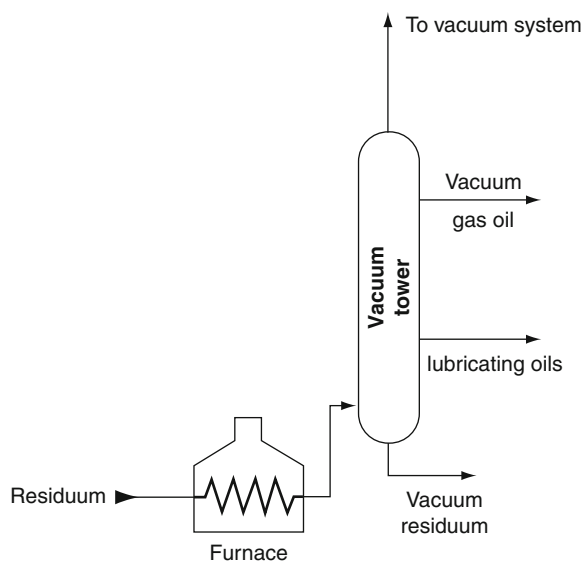
Petroleum Refining and Environmental Control and Environmental Effects. Figure 3

An atmospheric distillation unit (OSHA technical manual, Section IV, Chapter 2: Petroleum Refining Processes, [http://www.osha.gov/dts/osta/otm/otm\\_iv/otm\\_iv\\_2.html](http://www.osha.gov/dts/osta/otm/otm_iv/otm_iv_2.html))

and overhead system, vacuum furnace and tower, and bottoms exchanger. Hydrogen chloride may be present in the preheat exchanger, tower top zones, and overheads. Wastewater may contain water-soluble sulfides in high concentrations and other water-soluble compounds such as ammonia, chlorides, phenol, mercaptans, etc., depending upon the crude feedstock and the

treatment chemicals. Safe work practices and/or the use of appropriate personal protective equipment may be needed for exposures to chemicals and other hazards such as heat and noise, and during sampling, inspection, maintenance, and turnaround activities.

The primary source of emissions is combustion of fuels in the crude preheat furnace and in boilers that



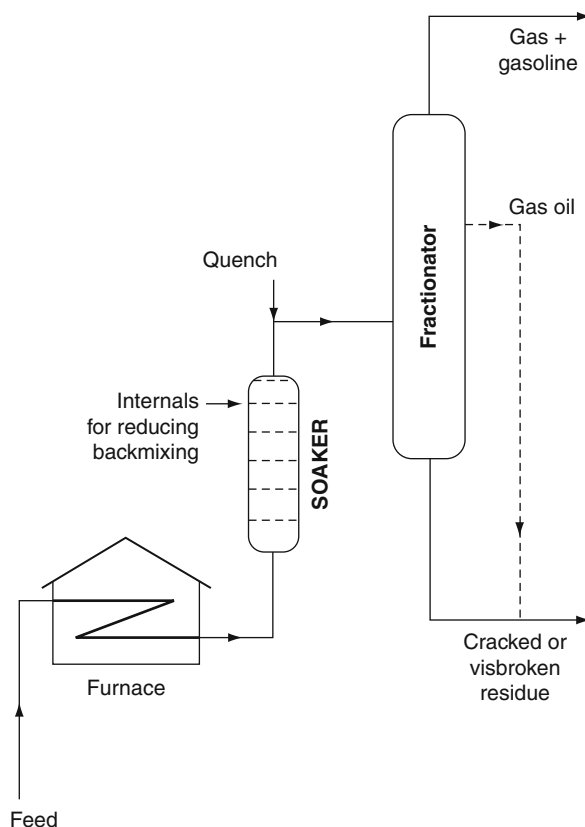
**Petroleum Refining and Environmental Control and Environmental Effects. Figure 4**

A vacuum distillation unit (OSHA technical manual, Section IV, Chapter 2: Petroleum Refining Processes, [http://www.osha.gov/dts/osta/otm/otm\\_iv/otm\\_iv\\_2.html](http://www.osha.gov/dts/osta/otm/otm_iv/otm_iv_2.html))

produce steam for process heat and stripping. When operating in an optimum condition and burning cleaner fuels (e.g., natural gas, refinery gas), these heating units create relatively low emissions of sulfur oxides, ( $\text{SO}_x$ ), nitrogen oxides ( $\text{NO}_x$ ), carbon monoxide (CO), hydrogen sulfide ( $\text{H}_2\text{S}$ ), particulate matter, and volatile hydrocarbons. If fired with lower grade fuels (e.g., refinery fuel pitch, coke) or operated inefficiently (incomplete combustion), heaters can be a significant source of emissions.

Petroleum distillation units generate considerable wastewater – often an oily sour wastewater and the constituents of sour wastewater streams include hydrogen sulfide, ammonia, suspended solids, chlorides, mercaptans, and phenol, characterized by a high pH.

**Visbreaking and Coking** *Visbreaking* (Fig. 5), like many thermal cracking processes, tends to produce a relatively small amount of fugitive emissions and sour wastewater [2, 6]. Usually some wastewater is produced from steam strippers and the fractionator. Wastewater is also generated during unit cleanup and cooling operations and from the steam injection

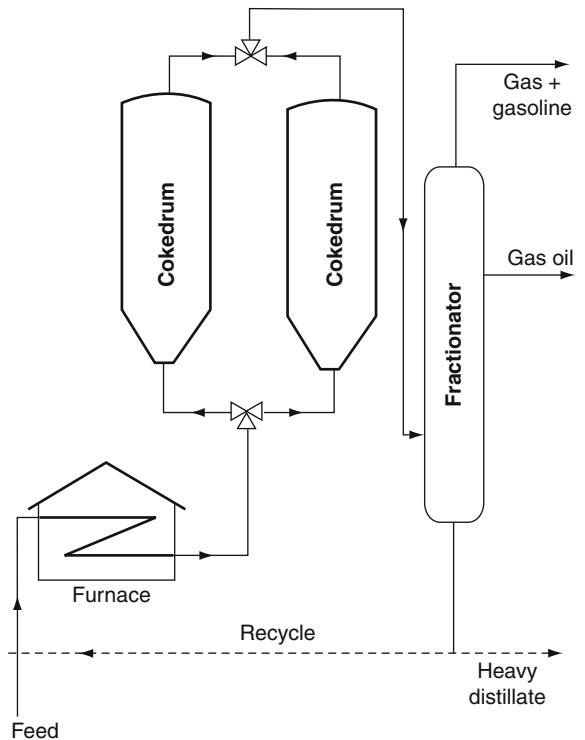


**Petroleum Refining and Environmental Control and Environmental Effects. Figure 5**

A soaker visbreaking unit (OSHA technical manual, Section IV, Chapter 2: Petroleum Refining Processes, [http://www.osha.gov/dts/osta/otm/otm\\_iv/otm\\_iv\\_2.html](http://www.osha.gov/dts/osta/otm/otm_iv/otm_iv_2.html))

process to remove organic deposits from the soaker or from the coil. Combined wastewater flows from thermal cracking and coking processes are about 3.0 gal per barrel of process feed.

*Delayed coking* is the oldest, most widely used process and has changed very little in the 5 or more decades in which it has been on stream in refineries [2]. *Fluid coking* is a continuous fluidized solids process that cracks feed thermally over heated coke particles in a reactor vessel to gas, liquid products, and coke [2]. Heat for the process is supplied by partial combustion of the coke, with the remaining coke being drawn as product. The new coke is deposited in a thin fresh layer on the outside surface of the circulating coke particle.



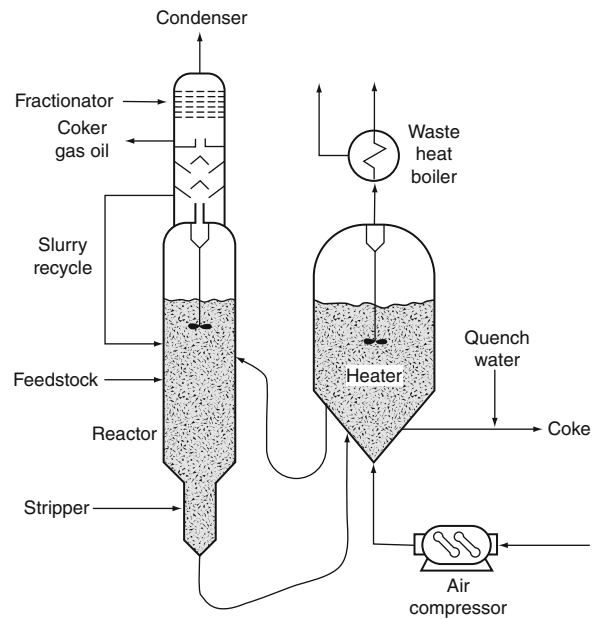
**Petroleum Refining and Environmental Control and Environmental Effects. Figure 6**

A delayed coking unit (OSHA technical manual, Section IV, Chapter 2: Petroleum Refining Processes, [http://www.osha.gov/dts/osta/otm/otm\\_iv/otm\\_iv\\_2.html](http://www.osha.gov/dts/osta/otm/otm_iv/otm_iv_2.html))

*Coking processes* (Figs. 6 and 7) produce a relatively small amount of sour wastewater from steam strippers and fractionators. Wastewater is generated during coke removal and cooling operations and from the steam injection process to cut coke from the coke drums. Combined wastewater flows from thermal cracking and coking processes are about 3.0 gal per barrel of process feed.

Particulate emissions from decoking can also be considerable. Coke-laden water from decoking operations in delayed cokers (hydrogen sulfide, ammonia, suspended solids), coke dust (carbon particles and hydrocarbons) occur.

**Fluid Catalytic Cracking** Fluid catalytic cracking (Fig. 8) is one of the largest sources of air emission in refineries [2, 6]. Air emissions are released in process

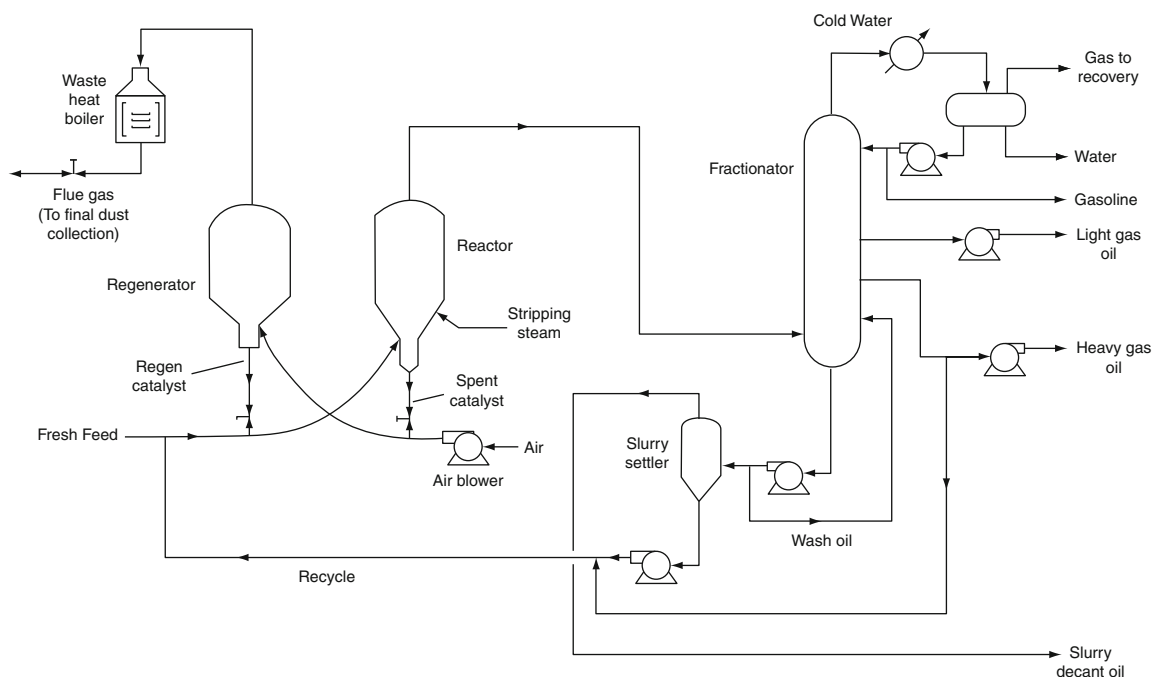


**Petroleum Refining and Environmental Control and Environmental Effects. Figure 7**

A fluid coking unit [12]

heater flue gas, as fugitive emissions from leaking valves and pipes, and during regeneration of the cracking catalyst. If not controlled, catalytic cracking is one of the most substantial sources of carbon monoxide and particulate emissions in the refinery. In non-attainment areas where carbon monoxide and particulates are above acceptable levels, carbon monoxide waste heat boilers (CO boiler) and particulate controls are employed. Carbon monoxide produced during regeneration of the catalyst is converted to carbon dioxide either in the regenerator or further downstream in a carbon monoxide waste heat boiler (CO boiler). Catalytic crackers are also significant sources of sulfur oxides and nitrogen oxides. The nitrogen oxides produced by catalytic crackers is expected to be a major target of emissions reduction in the future.

Catalytic cracking units, like coking units, usually include some form of fractionation or steam stripping as part of the process configuration. These units all produce sour waters and sour gases containing some hydrogen sulfide and ammonia. Like crude oil distillation, some of the toxic releases reported by the refining



**Petroleum Refining and Environmental Control and Environmental Effects. Figure 8**  
Schematic of a fluid catalytic cracking unit [12]

industry are generated through sour water and gases, notably ammonia. Gaseous ammonia often leaves fractionating and treating processes in the sour gas along with hydrogen sulfide and fuel gases [6].

Catalytic cracking (primarily fluid catalytic cracking) generates considerable sour wastewater from fractionators used for product separation, from steam strippers used to strip oil from catalysts, and in some cases from scrubber water. The steam stripping process used to purge and regenerate the catalysts can contain metal impurities from the feed in addition to oil and other contaminants. Sour wastewater from the fractionator/gas concentration units and steam strippers contains oil, suspended solids, phenols, cyanides, hydrogen sulfide, ammonia, spent catalysts, metals from crude oil, and hydrocarbons.

Catalytic cracking generates significant quantities of spent process catalysts (containing metals from crude oils and hydrocarbons) that are often sent off-site for disposal or recovery or recycling. Management options can include land filling, treatment, or separation and recovery of the metals. Metals deposited

on catalysts are often recovered by third-party recovery facilities. Spent catalyst fines (containing aluminum silicate and metals) from electrostatic precipitators are also sent off-site for disposal and/or recovery options.

Catalytic crackers also produce a significant amount of fine catalyst dust that results from the constant movement of catalyst grains against each other. This dust contains primarily alumina ( $\text{Al}_2\text{O}_3$ ) and small amounts of nickel (Ni) and vanadium (V), and is generally carried along with the carbon monoxide stream to the carbon monoxide waste heat boiler. The dust is separated from the carbon dioxide stream exiting the boiler through the use of cyclones, flue gas scrubbing, or electrostatic precipitators.

**Hydrocracking and Hydrotreating** *Hydrocracking* (Fig. 9) generates air emissions through process heater flue gas, vents, and fugitive emissions [2, 6]. Unlike fluid catalytic cracking catalysts, hydrocracking catalysts are usually regenerated off-site after months or years of operations, and little or no emissions or dust is

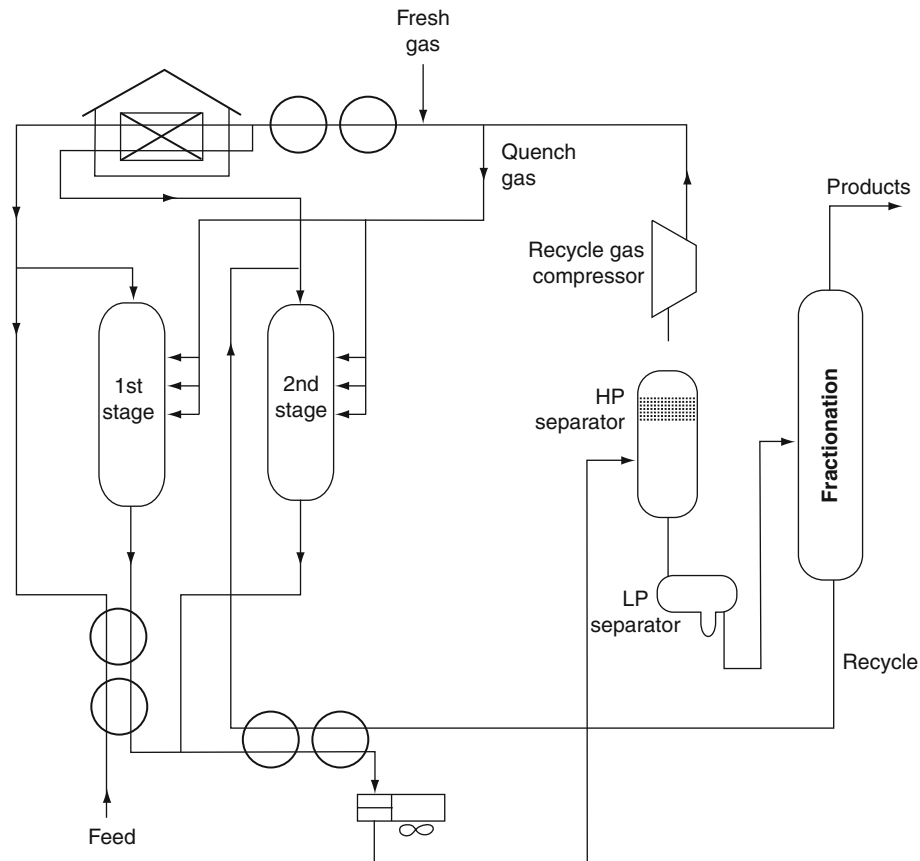
generated. However, the use of heavy oil as feedstock to the unit can change this balance.

Hydrocracking produces less sour wastewater than catalytic cracking. Hydrocracking, like catalytic cracking, produces sour wastewater at the fractionator. These processes include processing in a separator (API separator, corrugated plate interceptor) that creates sludge [2, 6]. Physical or chemical methods are then used to separate the remaining emulsified oils from the wastewater. Treated wastewater may be discharged to public wastewater treatment, to a refinery secondary treatment plant for ultimate discharge to public wastewater treatment, or may be recycled and used as process water. The separation process permits recovery of usable oil, and also creates

a sludge that may be recycled or treated as a hazardous waste.

Like catalytic cracking, hydrocracking processes generate toxic metal compounds, many of which are present in spent catalyst sludge and catalyst fines generated from catalytic cracking and hydrocracking. These include metals such as nickel (Ni), cobalt (Co), and molybdenum (Mo).

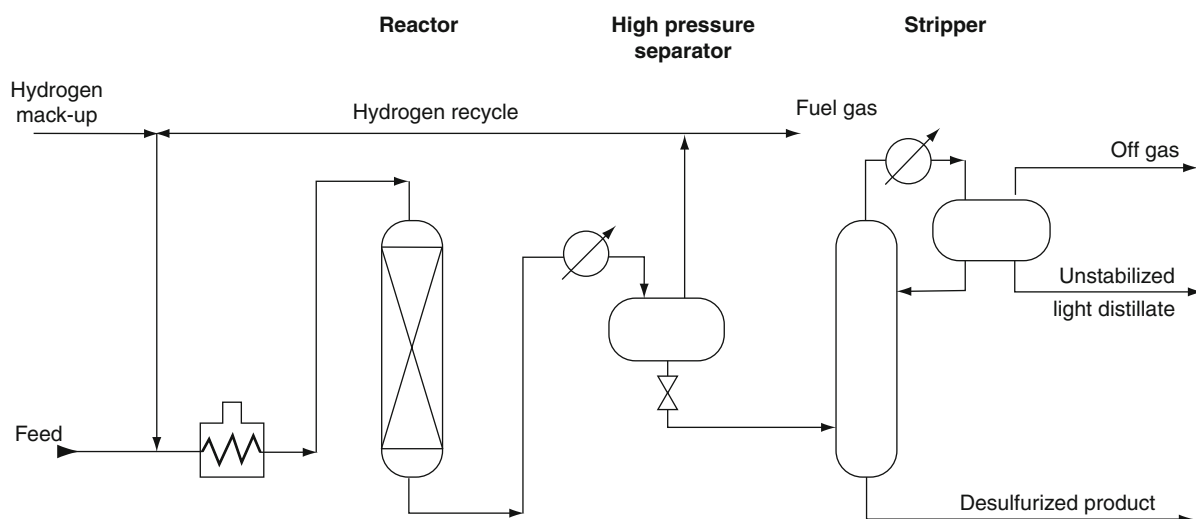
*Hydrotreating* is the less severe removal of heteroatomic species by treatment of a feedstock or product in the presence of hydrogen [2, 6]. The process (Fig. 10) generates air emissions through process heater flue gas, vents, and fugitive emissions [6]. Unlike fluid catalytic cracking catalysts, hydrotreating catalysts are usually regenerated off-site after months or years of



**Petroleum Refining and Environmental Control and Environmental Effects. Figure 9**

A two-stage hydrocracking unit (OSHA technical manual, Section IV, Chapter 2: Petroleum Refining Processes, [http://www.osha.gov/dts/osta/otm/otm\\_iv/otm\\_iv\\_2.html](http://www.osha.gov/dts/osta/otm/otm_iv/otm_iv_2.html))





**Petroleum Refining and Environmental Control and Environmental Effects. Figure 10**

A distillate hydrotreating unit (OSHA technical manual, Section IV, Chapter 2: Petroleum Refining Processes, [http://www.osha.gov/dts/osta/otm/otm\\_iv/otm\\_iv\\_2.html](http://www.osha.gov/dts/osta/otm/otm_iv/otm_iv_2.html))

operations, and little or no emissions or dust is generated from the catalyst regeneration process at the refinery. Air emissions factors for emissions from process heaters and boilers used throughout the refinery can be calculated (Emissions Factors & AP 42, *Compilation of Air Pollutant Emission Factors*).

Fugitive air emissions of volatile components released during hydrotreating may also be toxic components. These include toluene, benzene, xylenes, and other volatiles that are reported as toxic chemical releases under the EPA Toxics Release Inventory.

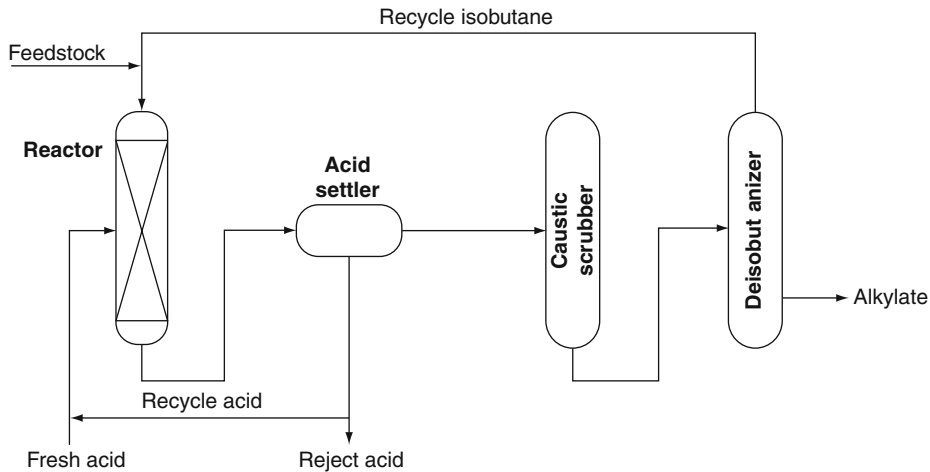
Hydrotreating generates sour wastewater from fractionators used for product separation. Like most separation processes in the refinery, the process water used in fractionators often comes in direct contact with oil, and can be highly contaminated. It also contains hydrogen sulfide and ammonia and must be treated along with other refinery sour waters.

Oily sludge from the wastewater treatment facility that result from treating oily and/or sour wastewaters from hydrotreating and other refinery processes may be hazardous wastes, depending on how they are managed. These include API separator sludge, primary treatment sludge, sludge from various gravitational separation units, and float from dissolved air flotation units.

Hydrotreating also produces some residuals in the form of spent catalyst fines, usually consisting of aluminum silicate and some metals (e.g., cobalt, molybdenum, nickel, tungsten). Spent hydrotreating catalyst is now listed as a hazardous waste (K171) (except for most support material). Hazardous constituents of this waste include benzene and arsenia (arsenic oxide,  $As_2O_3$ ). The support material for these catalysts is usually an inert ceramic (e.g., alumina,  $Al_2O_3$ ).

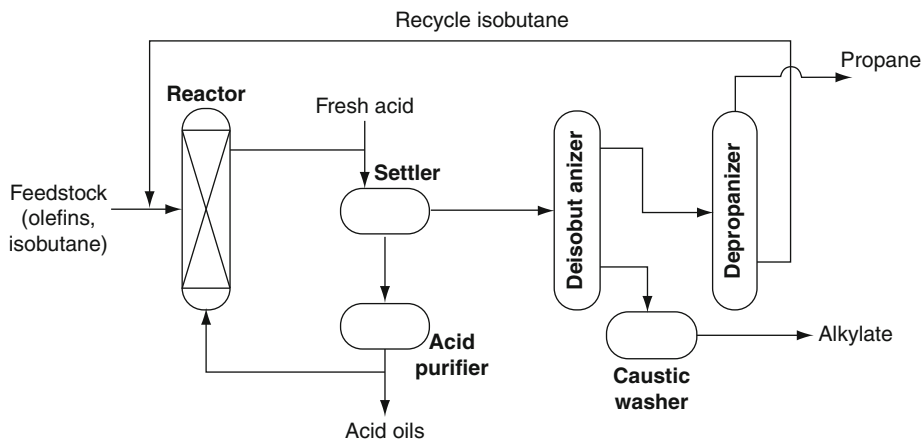
**Alkylation and Polymerization** Alkylation (Fig. 11) combines low-molecular-weight olefins (primarily a mixture of propylene and butylene) with isobutene in the presence of a catalyst, either sulfuric acid or hydrofluoric acid [2]. The product is called alkylate and is composed of a mixture of high-octane, branched-chain paraffinic hydrocarbons. Alkylate is a premium blending stock because it has exceptional antiknock properties and is clean burning. The octane number of alkylate depends mainly upon the kind of olefins used and upon operating conditions.

Emissions from alkylation processes (Figs. 11 and 12) and polymerization processes (Fig. 13) include fugitive emissions of volatile constituents in the feed, and



**Petroleum Refining and Environmental Control and Environmental Effects. Figure 11**

An alkylation unit (sulfuric acid catalyst) (OSHA technical manual, Section IV, Chapter 2: Petroleum Refining Processes, [http://www.osha.gov/dts/osta/otm/otm\\_iv/otm\\_iv\\_2.html](http://www.osha.gov/dts/osta/otm/otm_iv/otm_iv_2.html))



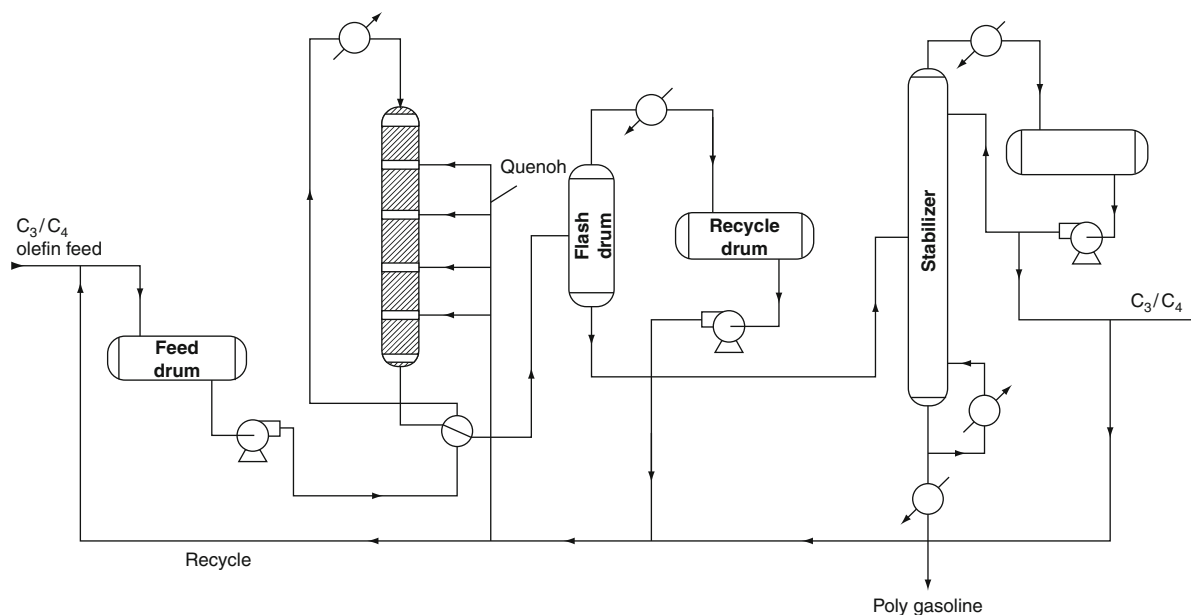
**Petroleum Refining and Environmental Control and Environmental Effects. Figure 12**

An alkylation unit (hydrogen fluoride catalyst) (OSHA technical manual, Section IV, Chapter 2: Petroleum Refining Processes, [http://www.osha.gov/dts/osta/otm/otm\\_iv/otm\\_iv\\_2.html](http://www.osha.gov/dts/osta/otm/otm_iv/otm_iv_2.html))

emissions that arise from process vents during processing. These can take the form of acidic hydrocarbon gases, nonacidic hydrocarbon gases, and fumes that may have a strong odor (from sulfonated organic compounds and organic acids, even at low concentrations). To prevent releases of hydrofluoric acid, refineries install a variety of mitigation and control technologies (e.g., acid inventory reduction, hydrogen fluoride detection systems, isolation valves, rapid acid transfer systems, and water spray systems).

In hydrofluoric acid alkylation processes, acidic hydrocarbon gases can originate anywhere hydrogen fluoride is present (e.g., during a unit upset, unit shutdown, or maintenance) [2, 6]. Hydrofluoric acid alkylation units are designed to pipe these gases from acid vents and valves to a separate closed-relief system where the acid is neutralized.

Another source of emissions is combustion of fuels in process boilers to produce steam for strippers. As with all process heaters in the refinery, these boilers



**Petroleum Refining and Environmental Control and Environmental Effects. Figure 13**

A polymerization process (OSHA technical manual, Section IV, Chapter 2: Petroleum Refining Processes, [http://www.osha.gov/dts/osta/otm/otm\\_iv/otm\\_iv\\_2.html](http://www.osha.gov/dts/osta/otm/otm_iv/otm_iv_2.html))

produce significant emissions of sulfur oxides, nitrogen oxides, carbon monoxide, particulate matter, and volatile hydrocarbons.

Alkylation generates relatively low volumes of wastewater, primarily from water washing of the liquid reactor products. Wastewater is also generated from steam strippers, depropanizers, and debutanizers, and can be contaminated with oil and other impurities. Liquid process waters (hydrocarbons and acid) originate from minor undesirable side reactions and from feed contaminants, and usually exit as a bottoms stream from the acid regeneration column. The bottom layer is an acid-water mixture that is sent to the neutralizing drum. The acid in this liquid eventually ends up as insoluble calcium fluoride.

Sulfuric acid alkylation generates considerable quantities of spent acid that must be removed and regenerated. Nearly all the spent acid generated at refineries is regenerated and recycled and, although technology for on-site regeneration of spent sulfuric acid is available, the supplier of the acid may perform this task off-site. If sulfuric acid production capacity is limited, acid regeneration is often done on-site. The development of internal acid regeneration for

hydrofluoric acid units has virtually eliminated the need for external regeneration, although most operations retain one for start-ups or during periods of high feed contamination.

Both sulfuric acid and hydrofluoric acid alkylation units generate neutralization sludge from treatment of acid-laden streams with caustic solutions in neutralization or wash systems. Sludge from hydrofluoric acid alkylation neutralization systems consists largely of calcium fluoride and unreacted lime, and is usually disposed of in a landfill. It can also be directed to steel manufacturing facilities, where the calcium fluoride can be used as a neutral flux to lower the slag-melting temperature and improve slag fluidity. Calcium fluoride can also be routed back to a hydrofluoric acid manufacturer.

A basic step in hydrofluoric acid manufacture is the reaction of sulfuric acid with fluorspar (calcium fluoride) to produce hydrogen fluoride and calcium sulfate. Spent alumina is also generated by the defluorination of some hydrofluoric acid alkylation products over alumina. It is disposed of or sent to the alumina supplier for recovery. Other solid residuals from hydrofluoric acid alkylation include any porous materials that may have come in contact with the hydrofluoric acid.

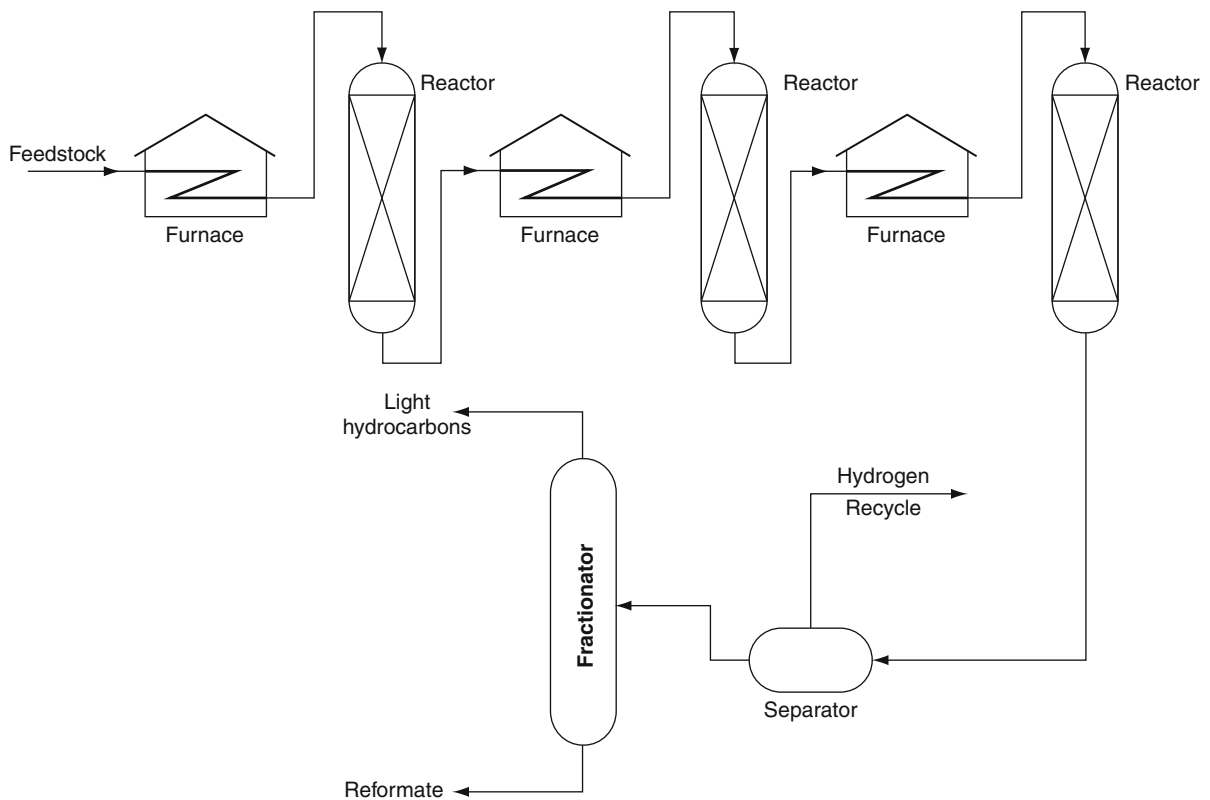
**Catalytic Reforming** Catalytic reforming (Fig. 14) converts alkanes to cycloalkanes and to aromatics and emissions from catalytic reforming include fugitive emissions of volatile constituents in the feed, and emissions from process heaters and boilers [2, 6]. As with all process heaters in the refinery, combustion of fossil fuels produces emissions of sulfur oxides, nitrogen oxides, carbon monoxide, particulate matter, and volatile hydrocarbons.

Benzene, toluene, and the xylene isomers are toxic aromatic chemicals that are produced during the catalytic reforming process and used as feedstocks in chemical manufacturing. Due to their highly volatile nature, fugitive emissions of these chemicals are a source of their release to the environment during the reforming process. Point air sources may also arise during the process of separating these chemicals.

In a continuous reformer, some particulate and dust matter can be generated as the catalyst moves

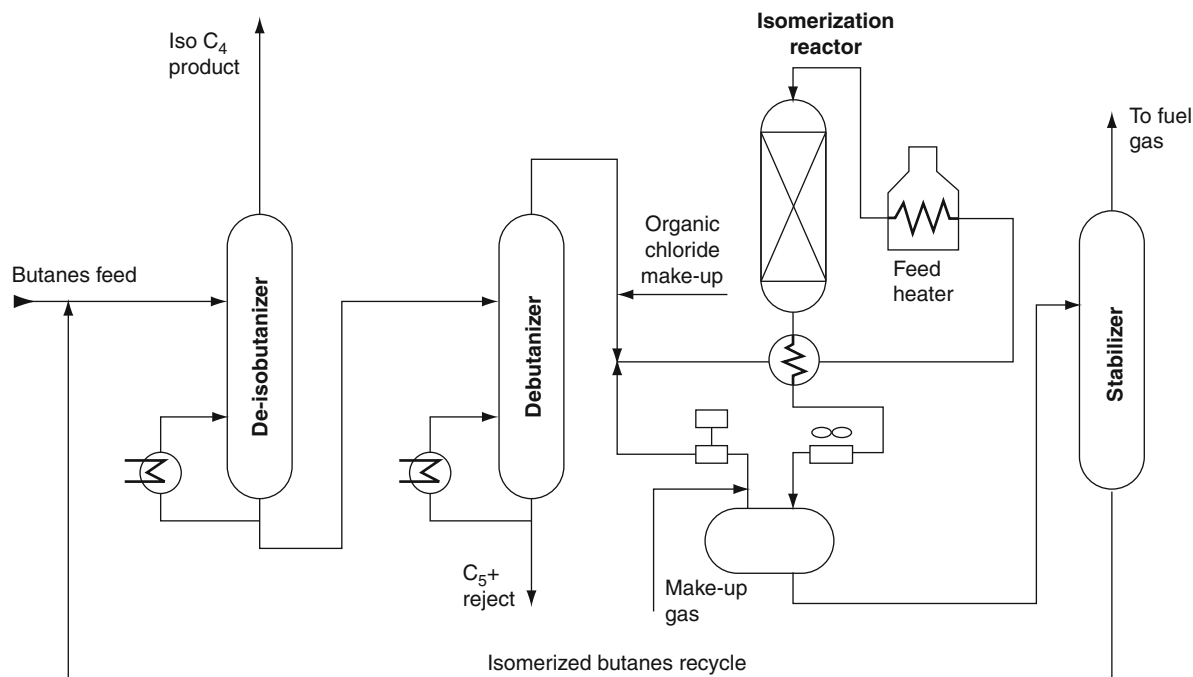
from reactor to reactor, and is subject to attrition. However, due to catalyst design, little attrition occurs, and the only outlet to the atmosphere is the regeneration vent, which is most often scrubbed with a caustic to prevent emission of hydrochloric acid (this also removes particulate matter). Emissions of carbon monoxide and hydrogen sulfide may occur during regeneration of catalyst.

**Isomerization** Isomerization (Fig. 15) converts n-butane, n-pentane, and n-hexane into their respective iso-paraffins of substantially higher octane number [2]. The straight-chain paraffins are converted to their branched-chain counterparts whose component atoms are the same but are arranged in a different geometric structure. Isomerization is important for the conversion of n-butane into iso-butane, to provide additional feedstock for alkylation units,



**Petroleum Refining and Environmental Control and Environmental Effects. Figure 14**

A catalytic reforming unit (OSHA technical manual, Section IV, Chapter 2: Petroleum Refining Processes, [http://www.osha.gov/dts/osta/otm/otm\\_iv/otm\\_iv\\_2.html](http://www.osha.gov/dts/osta/otm/otm_iv/otm_iv_2.html))



**Petroleum Refining and Environmental Control and Environmental Effects. Figure 15**

A butane isomerization unit (OSHA technical manual, Section IV, Chapter 2: Petroleum Refining Processes, [http://www.osha.gov/dts/osta/otm/otm\\_iv/otm\\_iv\\_2.html](http://www.osha.gov/dts/osta/otm/otm_iv/otm_iv_2.html))

and the conversion of normal pentanes and hexanes into higher branched isomers for gasoline blending.

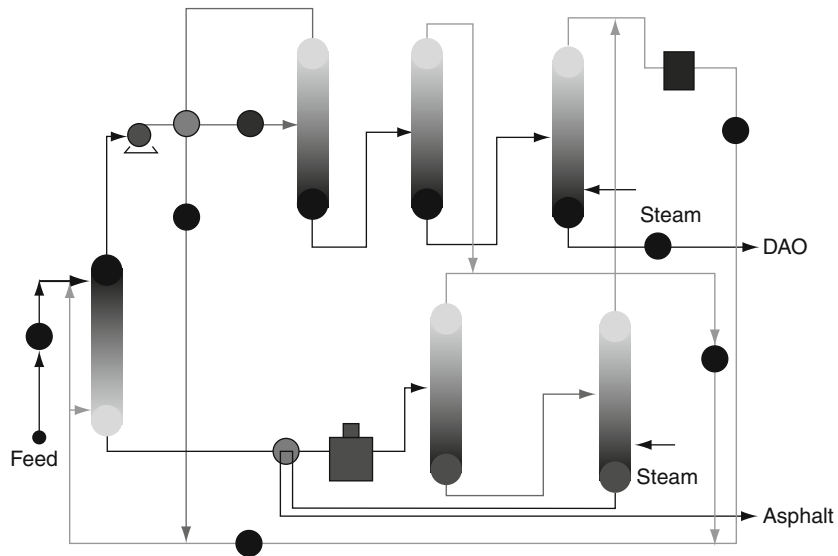
Isomerization processes produce sour water and caustic wastewater. The ether manufacturing process utilizes a water wash to extract methanol or ethanol from the reactor effluent stream. After the alcohol is separated, this water is recycled back to the system and is not released. In those cases where chloride catalyst activation agents are added, a caustic wash is used to neutralize any entrained hydrogen chloride. This process generates caustic wash water that must be treated before being released.

**Deasphalting and Dewaxing** Propane deasphalting (Fig. 16) produces lubricating oil base stocks by extracting asphaltenes and resins from vacuum distillation residua [2]. Propane is the usual solvent of choice due to its unique solvent properties. At lower temperatures (38–60°C, 100–140°F), paraffins are very soluble in propane, and at higher temperatures,

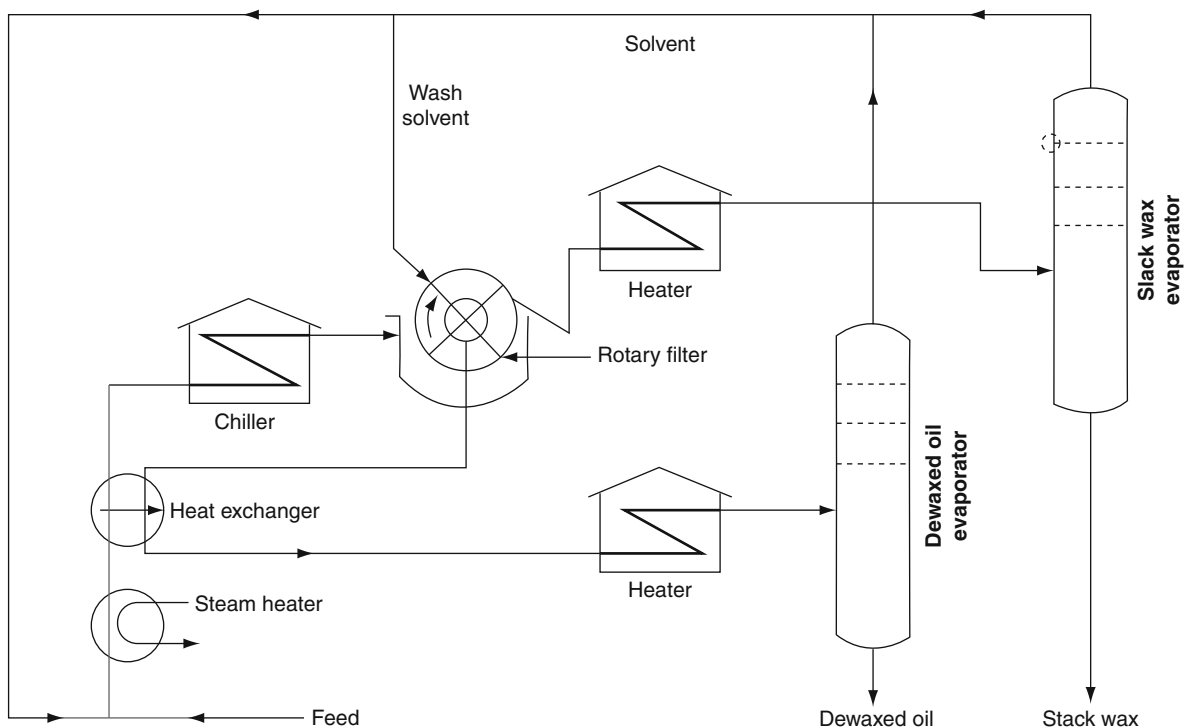
(approximately 93°C, 200°F) hydrocarbons are almost insoluble in propane. The propane deasphalting process is similar to solvent extraction in that a packed or baffled extraction tower or rotating disk contactor is used to mix the oil feed stocks with the solvent.

Air emissions may arise from fugitive propane emissions and process vents. These include heater stack gas (carbon monoxide, sulfur oxides, nitrogen oxides, and particulate matter) as well as hydrocarbon emission such as fugitive propane and fugitive solvents. Steam stripping wastewater (oil and solvents) and solvent recovery wastewater (oil and propane) are also produced.

Dewaxing (Fig. 17) processes also produce heater stack gas (carbon monoxide, sulfur oxides, nitrogen oxides, and particulate matter) as well as hydrocarbon emission such as fugitive propane and fugitive solvents [2, 6]. Steam stripping wastewater (oil and solvents) and solvent recovery wastewater (oil and propane) are also produced. The fugitive solvent emissions may be toxic (toluene, methyl ethyl ketone, methyl isobutyl ketone).



Petroleum Refining and Environmental Control and Environmental Effects. Figure 16  
A deasphalting unit



Petroleum Refining and Environmental Control and Environmental Effects. Figure 17

A solvent dewaxing unit (OSHA technical manual, Section IV, Chapter 2: Petroleum Refining Processes, [http://www.osha.gov/dts/osta/otm/otm\\_iv/otm\\_iv\\_2.html](http://www.osha.gov/dts/osta/otm/otm_iv/otm_iv_2.html))

## Gaseous Emissions

Gaseous emissions from petroleum refining create a number of environmental problems [2, 6]. During combustion, the combination of hydrocarbons, nitrogen oxide, and sunlight results in localized low levels of ozone, or smog. This is particularly evident in large urban areas and especially when air does not circulate well. Petroleum use in automobiles also contributes to the problem in many areas. The primary effects are on the health of those exposed to the ozone, but plant life has been observed to suffer as well.

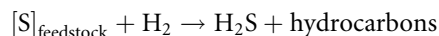
Refinery and natural gas streams may contain large amounts of acid gases, such as hydrogen sulfide (H<sub>2</sub>S) and carbon dioxide (CO<sub>2</sub>) [4, 5]. Hydrogen chloride (HCl), although not usually considered to be a major pollutant in petroleum refineries, can arise during processing from the presence of brine in petroleum that is incompletely dried. It can also be produced from mineral matter and other inorganic contaminants, gaining increasing recognition as a pollutant which needs serious attention.

Acid gases corrode refining equipment, harm catalysts, pollute the atmosphere, and prevent the use of hydrocarbon components in petrochemical manufacture. When the amount of hydrogen sulfide is large, it may be removed from a gas stream and converted to sulfur or sulfuric acid. Some natural gases contain sufficient carbon dioxide to warrant recovery as *dry ice*, i.e., solid carbon dioxide. And there is now a conscientious effort to mitigate the emission of pollutants from hydrotreating process by careful selection of process parameters and catalysts [2, 6, 15].

The terms “*refinery gas*” and “*process gas*” are also often used to include all of the gaseous products and by-products that emanate from a variety of refinery processes [5, 6]. There are also components of the gaseous products that must be removed prior to release of the gases to the atmosphere or prior to use of the gas in another part of the refinery, i.e., as a fuel gas or as a process feedstock.

Petroleum refining produces gas streams that often contain substantial amounts of acid gases such as hydrogen sulfide and carbon dioxide. More particularly

hydrogen sulfide arises from the hydrodesulfurization of feedstocks that contain organic sulfur:



Petroleum refining involves, with the exception of some of the more viscous crude oils, a primary distillation of the hydrogen mixture, which results in its separation into fractions differing in carbon number, volatility, specific gravity, and other characteristics [2, 6]. The most volatile fraction, that contains most of the gases which are generally dissolved in the crude, is referred to as *pipe still gas* or *pipe still light ends* and consists essentially of hydrocarbon gases ranging from methane to butane(s), or sometimes pentane(s).

The gas varies in composition and volume, depending on crude origin and on any additions to the crude made at the loading point. It is not uncommon to re-inject light hydrocarbons such as propane and butane into the crude before dispatch by tanker or pipeline. This results in a higher vapor pressure of the crude, but it allows one to increase the quantity of light products obtained at the refinery. Since light ends in most petroleum markets command a premium, while in the oil field itself propane and butane may have to be re-injected or flared, the practice of *spiking* crude oil with liquefied petroleum gas is becoming fairly common.

In addition to the gases obtained by distillation of petroleum, more highly volatile products result from the subsequent processing of naphtha and middle distillate to produce gasoline. Hydrogen sulfide is produced in the desulfurization processes involving hydrogen treatment of naphtha, distillate, and residual fuel, and from the coking or similar thermal treatments of vacuum gas oils and residual fuels. The most common processing step in the production of gasoline is the catalytic reforming of hydrocarbon fractions in the heptane (C<sub>7</sub>) to decane (C<sub>10</sub>) range.

In a series of processes commercialized under the generic name *reforming*, paraffin and naphthene (cyclic non-aromatic) hydrocarbons are altered structurally in the presence of hydrogen and a catalyst into aromatics, or isomerized to more highly branched hydrocarbons. Catalytic reforming processes thus not only result in the formation of a liquid product of higher octane

number, but also produce substantial quantities of gases. The latter are rich in hydrogen, but also contain hydrocarbons from methane to butanes, with a preponderance of propane ( $\text{CH}_3\text{CH}_2\text{CH}_3$ ), *n*-butane ( $\text{CH}_3\text{CH}_2\text{CH}_2\text{CH}_3$ ), and *iso*-butane [ $(\text{CH}_3)_3\text{CH}$ ].

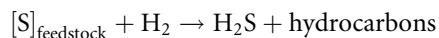
A second group of refining operations that contributes to gas production is that of the *catalytic cracking* processes [2, 6]. These consist of fluid-bed catalytic cracking in which heavy gas oils are converted into gas, liquefied petroleum gas, catalytic naphtha, fuel oil, and coke by contacting the heavy hydrocarbon with the hot catalyst. Both catalytic and thermal cracking processes, the latter being now largely used for the production of chemical raw materials, result in the formation of unsaturated hydrocarbons, particularly ethylene ( $\text{CH}_2=\text{CH}_2$ ), but also propylene (propene,  $\text{CH}_3\text{CH}=\text{CH}_2$ ), *iso*-butylene [*iso*-butene,  $(\text{CH}_3)_2\text{C}=\text{CH}_2$ ], and the *n*-butenes ( $\text{CH}_3\text{CH}_2\text{CH}=\text{CH}_2$ , and  $\text{CH}_3\text{CH}=\text{CHCH}_3$ ) in addition to hydrogen ( $\text{H}_2$ ), methane ( $\text{CH}_4$ ) and smaller quantities of ethane ( $\text{CH}_3\text{CH}_3$ ), propane ( $\text{CH}_3\text{CH}_2\text{CH}_3$ ), and butanes [ $\text{CH}_3\text{CH}_2\text{CH}_2\text{CH}_3$ ,  $(\text{CH}_3)_3\text{CH}$ ]. Diolefins such as butadiene ( $\text{CH}_2=\text{CH}\cdot\text{CH}=\text{CH}_2$ ) are also present.

Additional gases are produced in refineries with visbreaking and/or coking facilities that are used to process the heaviest crude fractions. In the visbreaking process, fuel oil is passed through externally fired tubes and undergoes liquid phase cracking reactions, which result in the formation of lighter fuel oil components. Oil viscosity is thereby reduced, and some gases, mainly hydrogen, methane, and ethane, are formed. Substantial quantities of both gas and carbon are also formed in coking (both delayed coking and fluid coking) in addition to the middle distillate and naphtha. When coking a residual fuel oil or heavy gas oil, the feedstock is preheated and contacted with hot carbon (coke) which causes extensive cracking of the feedstock constituents of higher molecular weight to produce lower molecular weight products ranging from methane, via liquefied petroleum gas and naphtha, to gas oil and heating oil. Products from coking processes tend to be unsaturated, and olefin components predominate in the tail gases from coking processes.

A further source of refinery gas is hydrocracking, a catalytic high-pressure pyrolysis process in the presence of fresh and recycled hydrogen. The

feedstock is again heavy gas oil or residual fuel oil, and the process is directed mainly at the production of additional middle distillates and gasoline. Since hydrogen is to be recycled, the gases produced in this process again have to be separated into lighter and heavier streams; any surplus recycle gas and the liquefied petroleum gas from the hydrocracking process are both saturated.

Both hydrocracker gases and catalytic reformer gases are commonly used in catalytic desulfurization processes. In the latter, feedstocks ranging from light to vacuum gas oils are passed at pressures of 500–1,000 psi with hydrogen over a hydrofining catalyst. This results mainly in the conversion of organic sulfur compounds to hydrogen sulfide,



The reaction also produces some light hydrocarbons by hydrocracking.

Thus refinery streams, while ostensibly being hydrocarbon in nature, may contain large amounts of acid gases such as hydrogen sulfide and carbon dioxide. Most commercial plants employ hydrogenation to convert organic sulfur compounds into hydrogen sulfide. Hydrogenation is effected by means of recycled hydrogen-containing gases or external hydrogen over a nickel molybdate or cobalt molybdate catalyst.

In summary, refinery process gas, in addition to hydrocarbons, may contain other contaminants, such as carbon oxides ( $\text{CO}_x$ , where  $x = 1$  and/or  $2$ ), sulfur oxides ( $\text{SO}_x$ , where  $x = 2$  and/or  $3$ ), as well as ammonia ( $\text{NH}_3$ ), mercaptans ( $\text{R-SH}$ ), and carbonyl sulfide ( $\text{COS}$ ).

The presence of these impurities may eliminate some of the sweetening processes, since some processes remove large amounts of acid gas but not to a sufficiently low concentration. On the other hand, there are those processes not designed to remove (or incapable of removing) large amounts of acid gases whereas they are capable of removing the acid gas impurities to very low levels when the acid gases are present only in low-to-medium concentration in the gas.

From an environmental viewpoint, not only are the means by which these gases can be utilized important but also equally important are the effects of these gases on the environment when they are introduced into the atmosphere.



In addition to the corrosion of equipment of acid gases, the escape into the atmosphere of sulfur-containing gases can eventually lead to the formation of the constituents of acid rain, i.e., the oxides of sulfur ( $\text{SO}_2$  and  $\text{SO}_3$ ). Similarly, the nitrogen-containing gases can also lead to nitrous and nitric acids (through the formation of the oxides  $\text{NO}_x$ , where  $x = 1$  or  $2$ ) which are the other major contributors to acid rain. The release of carbon dioxide and hydrocarbons as constituents of refinery effluents can also influence the behavior and integrity of the ozone layer.

Hydrogen chloride, if produced during refining, quickly picks up moisture in the atmosphere to form droplets of hydrochloric acid and, like sulfur dioxide, is a contributor to acid rain [6]. However, hydrogen chloride may exert severe local effects because, unlike sulfur dioxide, it does not need to participate in any further chemical reaction to become an acid. Under atmospheric conditions that favor a buildup of stack emissions in the area of a large industrial complex or power plant, the amount of hydrochloric acid in rainwater could be quite high.

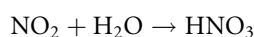
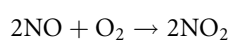
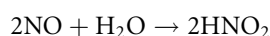
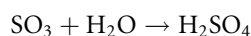
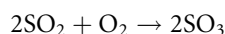
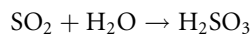
Natural gas is also capable of producing emissions that are detrimental to the environment. While the major constituent of natural gas is methane, there are components such as carbon dioxide ( $\text{CO}$ ), hydrogen sulfide ( $\text{H}_2\text{S}$ ), and mercaptans (thiols;  $\text{R-SH}$ ), as well as trace amounts of sundry other emissions. The fact that methane has a foreseen and valuable end-use makes it a desirable product, but in several other situations, it is considered a pollutant, having been identified a greenhouse gas.

A sulfur removal process must be very precise, since natural gas contains only a small quantity of sulfur-containing compounds that must be reduced several orders of magnitude. Most consumers of natural gas require less than 4 ppm in the gas.

A characteristic feature of natural gas that contains hydrogen sulfide is the presence of carbon dioxide (generally in the range of 1–4% v/v). In cases where the natural gas does not contain hydrogen sulfide, there may also be a relative lack of carbon dioxide.

*Acid rain* occurs when the oxides of nitrogen and sulfur that are released to the atmosphere during the combustion of fossil fuels are deposited (as soluble acids) with rainfall, usually at some location remote from the source of the emissions.

It is generally believed (the chemical thermodynamics are favorable) that acidic compounds are formed when sulfur dioxide and nitrogen oxide emissions are released from tall industrial stacks. Gases such as sulfur oxides (usually sulfur dioxide,  $\text{SO}_2$ ) as well as the nitrogen oxides ( $\text{NO}_x$ ) react with the water in the atmosphere to form acids:



Acid rain has a pH less than 5.0 and predominantly consists of sulfuric acid ( $\text{H}_2\text{SO}_4$ ) and nitric acid ( $\text{HNO}_3$ ). As a point of reference, in the absence of anthropogenic pollution sources, the average pH of rain is 6.0 (slightly acidic; neutral pH = 7.0). In summary, the sulfur dioxide that is produced during a variety of processes will react with oxygen and water in the atmosphere to yield environmentally detrimental sulfuric acid. Similarly, nitrogen oxides will also react to produce nitric acid.

Another acid gas, hydrogen chloride ( $\text{HCl}$ ), although not usually considered to be a major emission, is produced from mineral matter and the brines that often accompany petroleum during production and is gaining increasing recognition as a contributor to acid rain. However, hydrogen chloride may exert severe local effects because it does not need to participate in any further chemical reaction to become an acid. Under atmospheric conditions that favor a buildup of stack emissions in the areas where hydrogen chloride is produced, the amount of hydrochloric acid in rainwater could be quite high.

In addition to hydrogen sulfide and carbon dioxide, gas may contain other contaminants, such as mercaptans ( $\text{R-SH}$ ) and carbonyl sulfide ( $\text{COS}$ ). The presence of these impurities may eliminate some of the sweetening processes since some processes remove large amounts of acid gas but not to a sufficiently low concentration. On the other hand, there are those processes that are not designed to remove (or are incapable

of removing) large amounts of acid gases. However, these processes are also capable of removing the acid gas impurities to very low levels when the acid gases are there in low-to-medium concentrations in the gas.

On a regional level, the emission of sulfur oxides ( $\text{SO}_x$ ) and nitrogen oxides ( $\text{NO}_x$ ) can also cause the formation of acid species at high altitudes, which eventually precipitate in the form of *acid rain*, damaging plants, wildlife, and property. Most petroleum products are low in sulfur or are desulfurized, and while natural gas sometimes includes sulfur as a contaminant, it is typically removed at the production site.

At the global level, there is concern that the increased use of hydrocarbon-based fuels will ultimately raise the temperature of the planet (*global warming*), as carbon dioxide reflects the infrared or thermal emissions from the earth, preventing them from escaping into space (*greenhouse effect*). Whether or not the potential for global warming becomes real will depend upon how emissions into the atmosphere are handled. There is considerable discussion about the merits and demerits of the global warming theory [16] and the discussion is likely to continue for some time. Be that as it may, the atmosphere can only tolerate pollutants up to a limiting value. And that value needs to be determined. In the meantime, efforts must be made to curtail the use of noxious and foreign (non-indigenous) materials into the air.

In summary, and from an environmental viewpoint, petroleum and natural gas processing can result in similar, if not the same, gaseous emissions as coal [2, 4, 6]. It is a question of degree insofar as the composition of the gaseous emissions may vary from coal to petroleum but the constituents are, in the majority of cases, the same.

There are a variety of processes which are designed for sulfur dioxide removal from gas streams [2], but scrubbing process utilizing limestone ( $\text{CaCO}_3$ ) or lime [ $\text{Ca}(\text{OH})_2$ ] slurries have received more attention than other gas scrubbing processes.

The majority of the gas scrubbing processes are designed to remove sulfur dioxide from the gas streams; some processes show the potential for removal of nitrogen oxide(s).

## Liquid Effluents

It is convenient to divide the hydrocarbon components of petroleum into the following three classes:

- *Paraffins*: saturated hydrocarbons with straight or branched chains
- *Naphthenes (alicyclic hydrocarbons)*: saturated hydrocarbons containing one or more rings, each of which may have one or more paraffin side chains
- *Aromatic compounds*: hydrocarbons containing one or more aromatic nuclei (for example, benzene, naphthalene, and phenanthrene) which may be co-joined with (substituted) naphthene rings and/or paraffin side chains

Thermal processing can significantly increase the concentration of polynuclear aromatic hydrocarbons in the product liquid because the low-pressure hydrogen-deficient conditions favor aromatization of naphthene constituents and condensation of aromatics to form larger ring systems. To the extent that more compounds like benzo(a)pyrene are produced, the liquids from thermal processes will be more carcinogenic than asphalt.

The sludge produced on acid treatment of petroleum distillates [2, 6], even gasoline and kerosene, is complex in nature. Esters and alcohols are present from reactions with olefins; sulfonation products from reactions with aromatic compounds, naphthene compounds, and phenols; and salts from reactions with nitrogen bases. To these constituents must be added the various products of oxidation–reduction reactions: coagulated resins, soluble hydrocarbons, water, and free acid.

The disposal of the sludge is a comparatively simple process for the sludge resulting from treating gasoline and kerosene, the so-called light oils. The insoluble oil phase separates out as a mobile tar, which can be mixed and burned without too much difficulty.

In all cases, careful separation of reaction products is important to the recovery of well-refined materials. This may not be easy if the temperature has risen as a consequence of chemical reaction. This will result in a persistent dark color traceable to reaction products that are redistributed as colloids. Separation may also be difficult at low temperature because of high viscosity of the stock, but this problem can be overcome by dilution with light naphtha or with propane.

In addition, delayed coking also requires the use of large volumes of water for hydraulic cleaning of the coke drum. However, the process water can be recycled if the oil is removed by skimming and suspended coke particles are removed by filtration. If this water is used in a closed cycle, the impact of delayed coking on water treatment facilities and the environment is minimized. The flexicoking process offers one alternative to direct combustion of coke for process fuel. The gasification section is used to process excess coke to mixture of carbon monoxide (CO), carbon dioxide (CO<sub>2</sub>), hydrogen (H<sub>2</sub>), and hydrogen sulfide (H<sub>2</sub>S) followed by treatment to remove the hydrogen sulfide. Currently, maximum residue conversion with minimum coke production is favored over gasification of coke.

### Solid Effluents

Catalyst disposal is a major concern in all refineries. In many cases, the catalysts are regenerated at the refinery for repeated use. Disposal of spent catalysts is usually part of an agreement with the catalysts manufacturer whereby the spent catalyst is returned for treatment and re-manufacture.

The formation of considerable quantities of coke in the *coking* processes is a cause for concern since it not only reduces the yield of liquid products but also initiates the necessity for disposal of the coke. Stockpiling to coke may be a partial answer unless the coke contains leachable materials that will endanger the ecosystem as a result of rain or snow melt.

In addition, the generation and emission of sulfur oxides (particularly sulfur dioxide) originates from the combustion of sulfur-containing coke as plant fuel. Sulfur dioxide (SO<sub>2</sub>) has a wide range of effects on health and on the environment. These effects vary from bronchial irritation upon short-term exposure to contributing to the acidification of lakes. Emissions of sulfur dioxide, therefore, are regulated in many countries.

### Future Directions

The petroleum refining industry includes integrated process operations that are engaged in refining crude petroleum into refined petroleum products, especially liquid fuels such as gasoline and diesel as well as

processes that produce raw materials for the petrochemical industry.

Over the past 4 decades, the refining industry has experienced significant changes in oil market dynamics, resource availability, and technological advancements. Advancements made in exploration, production, and refining technologies allow utilization of resources such as heavy oil and tar sand bitumen that were considered economically and technically unsuitable in the middle decades of the past century. Along with the many challenges, it is imperative for refiners to raise their operations to new levels of performance. Merely extending today's performance incrementally will fail to meet most company's performance goals.

Petroleum refining in the twenty-first century will continue to be shaped by the factors such as consolidation of oil companies, dramatic changes in market demand, customization of products, and a decrease in the API gravity and sulfur content of the petroleum feedstocks. In fact, in addition to a (hopeful but unlikely) plentiful supply of petroleum, the future of the refining industry will base on the following factors such as (1) increased operating costs or investments due to stringent environmental requirements for facilities and products, (2) accelerating globalization resulting in stronger international petroleum price scenarios. The effect of these factors is likely to reduce refinery profit margins further, and petroleum companies worldwide will need to make significant changes in their operation and structure to be competitive on global basis.

As global petroleum consumption increases and resources are depleted, the production of fuels and petrochemicals from residua, heavy oil, and tar sand bitumen will increase significantly. In fact, over the next decade, refineries will need to adapt to receiving heavier oils as well as a range of bio-feedstocks. It is conceivable that current refineries could not handle such a diverse slate of feedstocks without experiencing shutdowns and related problems.

As feedstocks to refineries change, there must be an accompanying change in refinery technology as petroleum feedstocks are becoming highly variable. At the same time, more stringent anti-pollution regulations are forcing greater restrictions on fuel specifications. There are fundamental limitations on how far current

processes can go in achieving proper control over feedstock behavior. This means a movement from conventional means of refining heavy feedstocks by using (typically) a coking process to more development and use of more innovative processes that will produce the maximum yields of liquid fuels (or other desired products) fuels from the feedstock.

Thus, the need for the development of upgrading processes continues in order to fulfill the product market demand as well as to satisfy environmental regulations. One area in particular, the need for residuum conversion, technology has emerged as result of declining residual fuel oil market and the necessity to upgrade crude oil residua beyond the capabilities of the visbreaking, coking, and low-severity hydrodesulfurization processes.

Technological advances are on the horizon for alternate sources of transportation fuels. For example, gas-to-liquids and biomass-to-liquids are just two of the concepts currently under development. However, the state of many of these technologies coupled with the associated infrastructure required to implement them leaves traditional refining of petroleum hydrocarbons for transportation fuels as the *modus operandi* for the foreseeable future, which for the purposes of this text is seen to be 50 years. The near future challenge for refiners will be how to harness new technologies to remain alive in a changing global marketplace.

It is imperative for refiners to raise their operations to new levels of performance. Merely extending current process performance incrementally will fail to meet most future performance goals. To do this, it will be necessary to reshape refining technology to be more adaptive to changing feedstocks and product demand and to explore the means by which the technology and methodology of refinery operations can be translated not only into increased profitability but also into survivability.

Furthermore, environmental regulations could either preclude unconventional production or, more likely, raise the cost significantly. If future US laws limited and/or taxed greenhouse gas emissions, these laws will lead to substantial increase in the costs of production of fuels from unconventional sources. In addition to increases in the volumes of carbon dioxide, restrictions on access to water also could prove costly, especially in the arid or semiarid western States.

In addition, environmental restrictions on land use could preclude unconventional oil production in some areas of the United States.

The general prognosis for reduction of emissions or emission cleanup is optimistic. It is considered likely that most of their environmental impact of petroleum refining can be substantially abated. A considerable investment in retrofitting or replacing existing facilities and equipment will be needed although a conscious goal must be to improve the efficiency with which petroleum is transformed and consumed.

## Bibliography

### Primary Literature

1. Ray DL, Guzzo L (1990) Trashing the planet: how science can help us deal with acid rain, depletion of the ozone, and nuclear waste (among other things). Regnery Gateway, Washington, DC
2. Speight JG (2007) The chemistry and technology of petroleum, 4th edn. CRC Press/Taylor & Francis, Boca Raton
3. Majumdar SB (1993) Regulatory requirements for hazardous materials. McGraw-Hill, New York
4. Speight JG (1993) Gas processing: environmental aspects and methods. Butterworth Heinemann, Oxford
5. Speight JG (1996) Environmental technology handbook. Taylor & Francis, Washington, DC
6. Speight JG (2005) Environmental analysis and technology for the refining industry. Wiley, Hoboken
7. Lipton S, Lynch J (1994) Handbook of health hazard control in the chemical process industry. Wiley, New York
8. Boyce A (1997) Introduction to environmental technology. Van Nostrand Reinhold, New York
9. Speight JG (2009) Enhanced recovery methods for heavy oil and tar sands. Gulf, Houston
10. Carson PA, Mumford CJ (1995) The safe handling of chemicals in industry. Wiley, New York
11. Renzoni A, Fossi MC, Lari L, Mattei N (1994) Contaminants in the environment: a multidisciplinary assessment of risks to man and other organisms. CRC Press, Boca Raton
12. Edwards JD (1995) Industrial wastewater treatment: a guidebook. CRC Press, Boca Raton
13. Thibodeaux LJ (1995) Environmental chemodynamics. Wiley, New York
14. Meyers RA (1997) Handbook of petroleum refining processes, 2nd edn. McGraw-Hill, New York
15. Occelli ML, Chianelli R (1996) Hydrotreating technology for pollution control. Marcel Dekker, New York
16. Hileman B (1996) Environmental hormone disruptors focus of major research initiatives. Chem Eng News 74(34):33

## Books and Reviews

- Abraham H (1945) *Asphalts and allied substances*, vol I. Van Nostrand, New York
- Forbes RJ (1958) *A history of technology*, vol V. Oxford University Press, Oxford
- Gary JH, Handwerk GE (2001) *Petroleum refining: technology and economics*, 4th edn. Marcel Dekker, New York
- Hoiberg AJ (1960) *Bituminous materials: asphalts, tars and pitches*, vol I & II. Interscience, New York
- Hsu CS, Robinson PR (2006) *Practical advances in petroleum processing*, vol 1 and 2. Springer, New York
- Khan MR, Patmore DJ (1997) Heavy oil upgrading processes. In: Speight JG (ed) *Petroleum chemistry and refining*. Taylor & Francis, Washington, DC. Chapter 6
- McKetta JJ (ed) (1992) *Petroleum processing handbook*. Marcel Dekker, New York
- Shih SS, Oballa MC (eds) (1991) *Tar sand upgrading technology*. Symposium series No. 282. American Institute for Chemical Engineers, New York
- Speight JG (2000) *The desulfurization of heavy oils and residua*, 2nd edn. Marcel Dekker, New York
- Speight JG, Ozum B (2002) *Petroleum refining processes*. Marcel Dekker, New York
- Speight JG (2008) *Handbook of synthetic fuels: properties, processes, and performance*. McGraw-Hill, New York

## PHEVs and BEVs in Coupled Power and Transportation Systems

MLADEN KEZUNOVIC<sup>1</sup>, S. TRAVIS WALLER<sup>2</sup>

<sup>1</sup>Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA

<sup>2</sup>Department of Civil Engineering, The University of Texas at Austin, Austin, USA

### Article Outline

Glossary  
 Definition  
 Introduction  
 Policy Issues  
 Modeling of Complex Systems  
 Benefits  
 Conclusions and Future Research  
 Acknowledgments  
 Bibliography

## Glossary

- BEV** Battery electric vehicle.
- DSM** Demand side Management; utility-sponsored programs to influence the time of use and amount of energy use by select customers.
- G2V** Grid-to-vehicle; using the electrical grid to charge the battery of a vehicle.
- HEV** Hybrid electric vehicle.
- OM** Outage management; set of manual and/or automated procedures used by operators of electric distribution systems to assist in restoration of power.
- PHEV** Plug-in hybrid electric vehicle.
- V2B** Vehicle-to-building; exporting electrical power from a vehicle battery into a building.
- V2G** Vehicle-to-grid; exporting electrical power from a vehicle battery to the electrical grid.

## Definition

With the price of oil peaking in the recent past close to the once unimaginable \$150 per barrel and the threat of global climate change increasingly acknowledged, the transportation sector is employing a number of new technologies that will enhance energy security by reducing the current dependency on oil-based fuels. Should the gasoline cost increase in the future, Plug-in Hybrid Electric Vehicles (PHEVs) and Battery Electric Vehicles (BEVs) will become the economical choice for transportation. Widespread adoption of PHEVs/BEVs will also improve air quality and carbon footprint, since point source pollution is easier to control than mobile source pollution. This level of control is essential for effective implementation of carbon cap-and-trade markets, which should spur further innovation. In USA, sales of Hybrid Electric Vehicles (HEVs) have grown 80% each year since 2000, proving that PHEVs/BEVs are likely an eventual reality that must be dealt with [1]. The implications of this reality will be highly dependent on the policies in place to use PHEVs/BEVs to the benefit of the transportation and power systems, as well as the drivers, industry, and public at large.

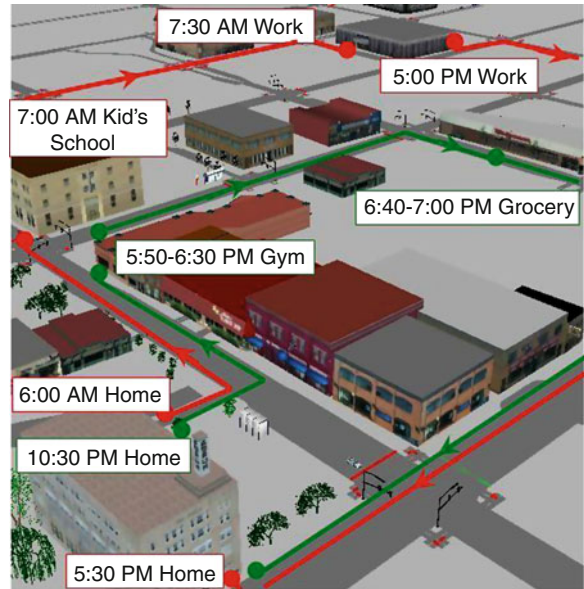
Beyond fuel costs and sustainability, the primary concern of the transportation sector is congestion. In 2005, congestion was estimated to cost the US economy \$78.2 billion in wasted time and fuel [2]. If PHEV/BEV drivers are given appropriate incentives

(e.g., strategically placed energy exchange stations), traveler behavior (e.g., choice of routing, departure time, and destination) impacting congestion may be affected.

In addition, the power industry is currently challenged to maintain reliability of operation while expanding the grid to meet growing demand. Large blackout such as the northeastern one in 2003 may create loses in billions of dollars [3]. Introducing the renewable resources to meet growing demand requires energy storage to deal with interfacing [4]. If proper policy is in place PHEVs/BEVs can provide a promising solution acting as mobile decentralized storage (MDS) of electrical energy. In this capacity, PHEVs/BEVs can serve in two modes: grid-to-vehicle (G2V) and vehicle-to-grid (V2G), each providing benefits to the power system operation. The G2V mode can be used to charge PHEVs/BEVs at reduced cost when the power system load is reduced and generation capacity is abundant, such as during night time. The V2G mode may be used when demand is high or supply is accidentally lost since the stored electric energy can be released from PHEVs/BEVs in an aggregated way, which will offer major contributions to regulation service and spinning reserves, as well as load-shedding prevention. The mobility of the energy storage in PHEVs/BEVs allows for strategic placement of the distributed generation source to optimize power system needs.

Figure 1 illustrates the spatial and temporal coupling of the power and transportation systems through showing an example of a PHEV/BEV driver's route, highlighting destinations where the driver could potentially engage in G2V or V2G activity. Options for meeting selected criteria for electricity and transportation networks simultaneously are numerous. Developing policy strategy requires understanding of trade-offs involved in pursuing certain solutions at the expense of others. The all-encompassing theoretical framework for such studies to the best of our knowledge is not available.

Traditionally, scientists have adopted a divide and conquer approach to understanding complex phenomena. Unfortunately, systems with emergent dynamics that are dominated by contextual interactions are not well suited to this classical approach (e.g., [5–7]). In such cases, directly addressing the couplings of system components may actually hasten progress. While this



**PHEVs and BEVs in Coupled Power and Transportation Systems. Figure 1**

Temporal and spatial dimensions of plug-in opportunities

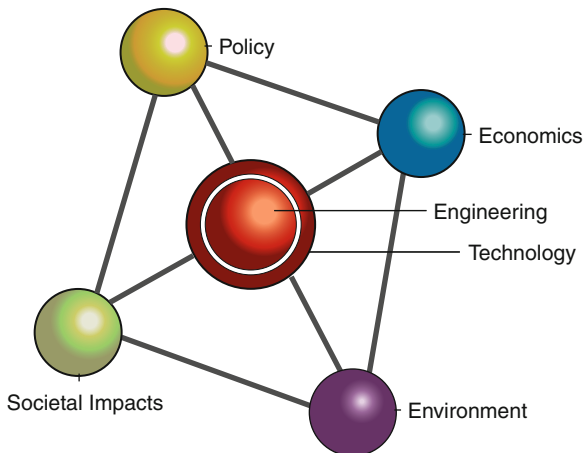
linkage presents new opportunities to improve the functioning and capacity utilization of each system, it also raises the spectrum of increasing dynamic complexity and cascading failures across systems.

In this entry, several open policies and research goals will be discussed, which facilitate optimizing the integration of the transportation systems and the behavior of its travelers with the electricity systems and behavior of its end-customers. PHEVs/BEVs based demand side management (DSM) and outage management (OM) are also presented as an application of PHEVs/BEVs using in the coupled power and transport system.

## Introduction

The impacts PHEVs/BEVs will have on transportation systems, power systems, and air quality are very complex. Studies conducted to date on this topic make many assumptions to simplify the problem. As stated in the definition, the problem space must be treated as one large complex system in order to capture emergent behavior.

The complexity of the issues involved in studying PHEVs/BEVs and their interaction with electricity and transportation networks is shown in Fig. 2, where



**PHEVs and BEVs in Coupled Power and Transportation Systems. Figure 2**

Illustration of multidisciplinary nature of problem

several disciplines that need to be involved in researching this multidisciplinary problem are shown.

Recent analyses confirm the feasibility of the grid-to-vehicle (G2V) and vehicle-to-grid (V2G) concepts [8–13]. The Electric Power Research Institute speculates that V2G could reduce the requirement for global, central-station generation capacity by up to 20% by the year 2050 [14]. Several studies omit any consideration of vehicle locations and desired activity patterns and assume a percentage of vehicles are plugged in and available when estimating the benefits to the grid and to drivers [8, 10, 11].

Many researchers have investigated the various potential benefits and implementation issues of the V2G concept. Kempton and Tomić studied the fundamentals of using PHEVs/BEVs for load leveling, regulation, reserve, and other purposes [15, 16]. Hadley and Tsvetkova analyzed the potential impacts of PHEVs on electricity demand, supply, generation, structure, prices, and associated emission levels in 2020 and 2030 in 13 regions specified by the North American Electric Reliability Corporation (NERC) [17]. Meliopoulos et al. considered the impacts of PHEVs/BEVs on electric power network components [18]. Anderson et al. performed the case studies of PHEVs/BEVs as regulating power providers in Sweden and Germany [19]. Guille and Gross presented a proposed framework to effectively integrate the aggregated battery electric vehicles into the grid as distributed energy

resources [20]. The combined impact PHEVs/BEVs make on both electric power system and transportation network has not been explored as much. When considering the role of PHEVs/BEVs as dynamically configurable (mobile) energy storage, the potential impacts on both electricity and transportation networks may become quite diverse. The flow of traffic is an important factor in deciding the flow of electric power that could be utilized from PHEVs/BEVs. Correlating the movement of people to the movement of the power load offers new opportunities in the smart grid.

One of the major advantages of PHEVs/BEVs is their usefulness as an MDS. MDS is a revolutionary concept because currently the power grid has no storage except for 2.2% of its capacity in pumped storage [11]. Without significant and reliable storage of energy, maintaining grid stability and reliability under the growing electricity demand is a complex problem. Utilities may contract with others to provide power in any one of the four types of markets: base-load power, peak power, spinning reserves, and regulation services. Several studies have shown that PHEVs/BEVs can provide ancillary services (spinning and regulation) at a profit [8, 10, 11]. Spinning reserves receive payment for providing continuous capacity regardless of whether energy is provided, and receive further payment if called on to feed energy into the grid. Regulation services feed a nominal amount of energy into the grid, and receive payment for reducing or increasing their energy consumption as needed. In the case of PHEVs/BEVs, being plugged-in in a predictable way means that capacity is available to feed into the system if called upon. PHEVs/BEVs are particularly well suited for regulation services since the impact on vehicle's energy resources may be zero.

The pricing of V2G and G2V services is expected to cause a fundamental shift in the behavior of PHEVs/BEVs drivers. Further research is needed to investigate the exact nature of this shift; however, if the pricing schemes are developed with both the power system and transportation system in mind, then PHEVs/BEVs could help solve problems plaguing the traffic network, particularly congestion. The pricing scheme should also consider air quality impacts caused by charging at different times in the day. As mentioned earlier, MDS will allow for renewable energy to be used more

efficiently. There will however remain times of the day more dominated by “dirty” fuels than others.

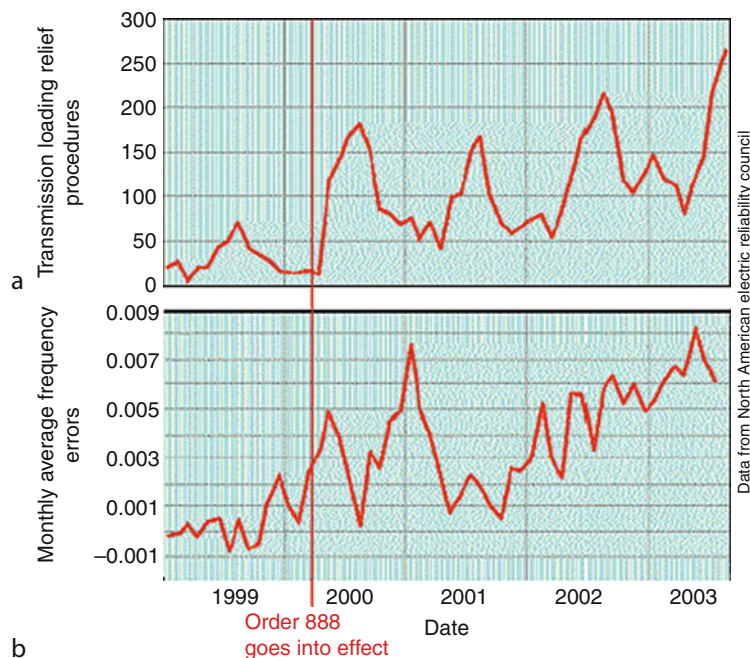
As observed, the body of research literature related to the multidimensional impact of PHEVs/BEVs is quite small. The remainder of this section will focus separately on the dual problems of improving the stability and reliability of the electrical grid and improving the efficiency of the roadway network.

### Stability and Reliability of the Electric Grid

Stability and reliability of the US electric grid have become issues of increasing concern since the occurrence of several blackouts in the 1990s (Western Interconnect in 1994 and 1996, and the Eastern Interconnect in 1999) and system deregulation. The devastating impact of the northeast blackout from August 14, 2003 reminded that the situation with the grid is only worsening and not improving. Here, a stable system is defined as one in which the phase and frequency of power generation units are constant. Ability of the system to maintain the state of equilibrium during normal and abnormal conditions is a measure of stability. Reliability is defined as the ability of the system

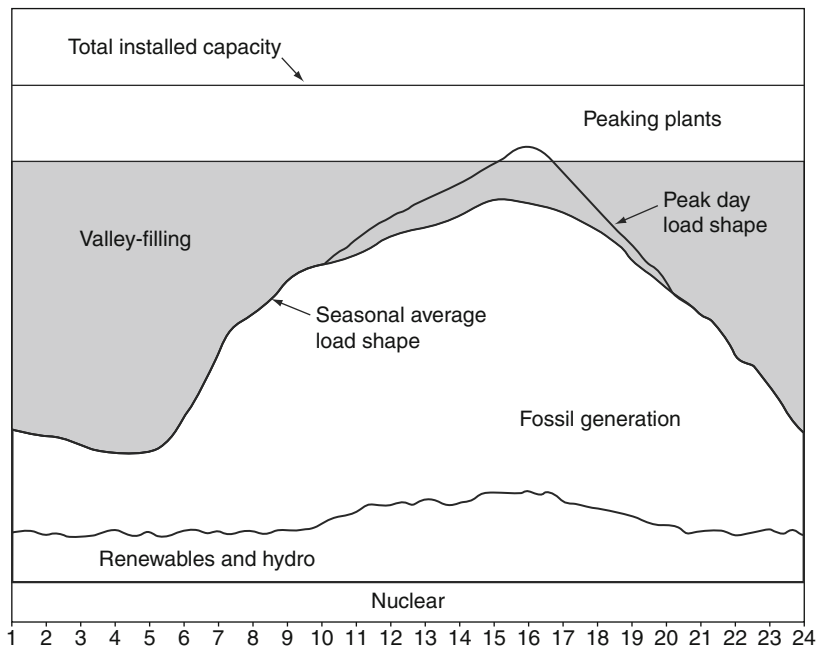
to meet unexpected demand and respond to failures. Ability of the system to deliver electricity to customers within the accepted standards, which may be affected by the failure rate, repair rate, or duration of loss, is a measure of reliability. Figure 3 illustrates the worsening stability problem. Order 888 in Fig. 3 relates to the Open Access to Transmission issued by Federal Regulatory Commission in 1996, which is the result of an authorization passed by the Congress as a part of the Energy Policy Act of 1992.

A major challenge in achieving these goals (stability and reliability) is the lack of energy storage. Figure 4 depicts the peaking structure of an example power load over the course of 1 day. In this example, demand grows rapidly starting at 6 a.m. and begins to decline after hitting a peak around 3 p.m. This peaking phenomenon is especially important to consider given that different energy sources are available at different times of day. For example, wind energy is most widely available at night when the demand for power is the lowest. While it may seem intuitive that a flat demand curve is the ideal, this is not necessarily true. More research is required to determine if parts of the system (e.g., transformers) require time to cool down. The large scale use



**PHEVs and BEVs in Coupled Power and Transportation Systems. Figure 3**  
Illustration of multidisciplinary nature of problem [21]





PHEVs and BEVs in Coupled Power and Transportation Systems. Figure 4

Illustrative peaking of electricity load [12]

of energy storage would significantly help meeting the stability and reliability needs, including managing the load variations shown in Fig. 4.

### Efficiency of the Roadway Network

Congestion is a problem not only in the electricity grid network, but also in the roadway network. Vehicle miles traveled (VMT) has risen consistently since the advent of the automobile, with dips when gasoline prices rise quickly (See Fig. 5 for the VMT trend since 1992). If the transportation sector is shifted to an alternative fuel source (i.e., electricity) with greater price stability, and especially if the source of the fuel is renewable, then VMT is expected to continue to increase into the foreseeable future. While mobility is an indicator of economic success, the expansion of a roadway system is limited by available space and finances. Roadway network efficiency is further constrained by the individual autonomy of drivers who act in their self-interest instead of the interest of the system (see [22] for a theoretical description of traveler behavior).

Extensive research has been conducted on improving the efficiency of the transportation system via

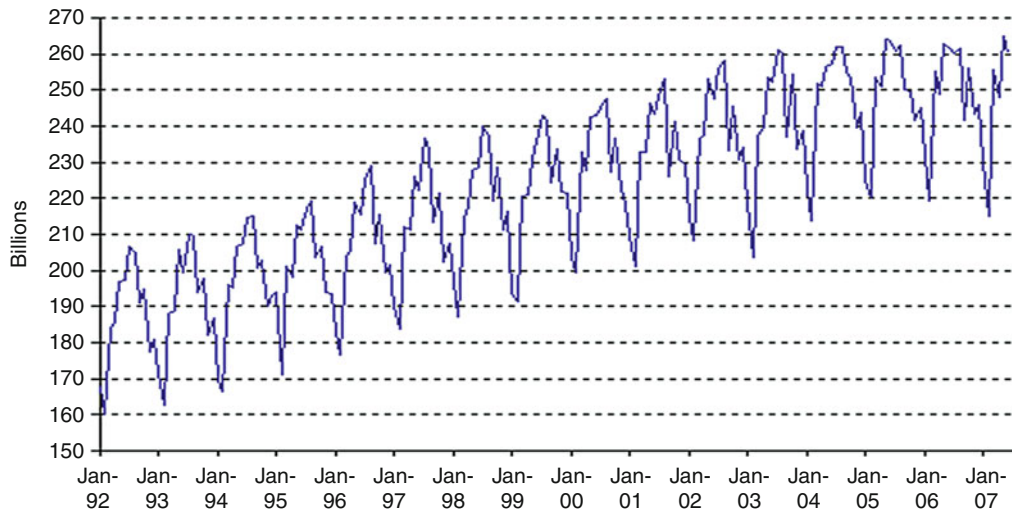
methods such as pricing and technology, but few solutions proposed offer a case even close to being as comprehensive as PHEVs/BEVs.

### Policy Issues

The policy issues presented here are centered on incentives to help industry develop and bring new value to end users of electricity and transportation networks, and society at large, while encouraging competition and development of new business opportunities.

### Improve Electric Grid Performance

Widespread deployment of PHEVs/BEVs will allow for increased energy storage, and improved reliability and stability of the electric grid. Linking the transportation and power systems through PHEVs/BEVs will allow for electrical energy storage on a scale much larger than is currently feasible. The additional energy capacity will be directly proportional to the penetration of PHEVs/BEVs into the automobile market, and modeling (see the modeling section) is needed to determine the exact increase in capacity across the space and time dimensions.



**PHEVs and BEVs in Coupled Power and Transportation Systems. Figure 5**  
US highway VMT [23]

The new mobile storage can only benefit the electric grid if it is available at the right time and place to service the grid when needed. To determine PHEVs/BEVs' demand for electric energy across space and time, travel patterns must be considered. Figure 1 shows an example of such a pattern, highlighting several destinations where a driver could potentially engage in G2V or V2G.

Stability and reliability issues were mentioned earlier. V2G is poised to greatly aid the grid in becoming more reliable and stable because vehicles are only in use for a small portion of each day (average daily travel time person in 2001 was 82.3 min [24]). During the remainder of the day, the vehicles can be plugged in and provide services (e.g., ancillary or regulatory).

This approach requires a policy shift to allow use of the MDS for energy to maintain stability and reliability. Also, policy that encourages utilities to cooperate with the PHEV/BEV owners or aggregators and provide tariff incentives for their participation in programs aimed at demand and distributed generation management and optimization is missing at the moment.

### **Enhance Penetration of Renewable Energy Sources to Improve Energy Security**

Increasing energy capacity by using PHEVs/BEVs as MDS will allow for increased investment in

renewable energies by alleviating concerns related to the temporarily highly variable nature of solar (daytime) and wind (primarily nighttime). Using renewable energy has benefits not only for the environment and air quality, but also for energy security by reducing reliance on the supply from oil producing countries.

This approach requires a policy shift to allow and encourage large scale use of the MDS for energy to support interfacing of renewable generation.

### **Reduce and Redistribute Pollution in the Electric Grid and Transportation Network**

By shifting the source of pollution away from vehicles, PHEVs/BEVs will change the transportation-based air pollution problem from a mobile source issue to a point source issue. This redistribution of pollution will likely have the effect of reducing pollution because point sources are much easier to control and some already have emission caps in place. Hadley [25] conducted initial research into the potential air quality impacts of PHEVs/BEVs, describing the impacts of G2V charging on air quality, and considering the types of power generation that are typically used at different times of day (e.g., coal-fired generation is prevalent at night in some regions).

Some policies are already in place to ensure that the redistribution of pollution that will occur with widespread deployment of PHEVs/BEVs will actually lead to a reduction in pollution. Further policy analysis is needed to ensure that V2G and G2V services are incentivized to occur at times when it will result in the maximum improvement in air quality.

### **Create New Markets and Further Deregulate Existing Markets**

PHEVs/BEVs are poised to open new markets and increase opportunities in existing ones. Carbon-trade markets should be aided because they facilitate the change of the transportation-based air pollution problem from a mobile source issue to a point source issue (as described earlier). While point sources of pollution are much easier to control, if they are nonrenewable, they will likely need to trade carbon credits to counter the increased emissions.

PHEVs/BEVs will also create new modes for participation in the electricity markets. There will be opportunities for businesses to act as Qualified Scheduling Entities (QSE) to the electric utility by facilitating V2G/G2V interactions. Such a QSE that aggregates across vehicles is necessary because any one vehicle's contribution will be too small to allow it to participate directly in the market. PHEVs/BEVs will likely function akin to small generators as a distributed energy resource.

Policy that enhances market development and deregulation allowing a new type of QSE to bid in a variety of markets is needed to facilitate the aggregated use of PHEVs/BEVs in "transportation-energy" markets.

### **Plan and Develop Energy Exchange Stations**

Energy exchange stations (for G2V and V2G) could take one of at least two forms. The first, the way considered by most electric vehicle research to date, assumes that individual drivers plug in and charge their vehicle over a period of several hours. Some examples of potential charging station locations are shopping malls, recreational areas, schools, and of course homes.

Further, rather than requiring drivers to plug into the grid and wait several hours to charge their batteries, battery exchange locations could be as ubiquitous as gas stations and automatically exchange discharged batteries with fully charged batteries. Charging PHEVs in this way has benefits for drivers because the process takes only a few minutes as opposed to several hours. Also, this system would require a leasing system for batteries similar to the system in place for leasing cell phones, alleviating driver concerns about battery life. The benefit for utilities is that control over charging and servicing the grid is centralized.

In reality, charging (G2V) and discharging (V2G) services will likely be based on a hybrid of the two methods mentioned above (individual drivers plugging into the grid and stations designed to exchange batteries). Depending on the pricing structure in place, it may make sense for drivers to exchange batteries during long drives and plug in to a household plug at night. Incentive structures will need to be developed that consider the different players – energy exchange stations and individual drivers.

The temporal and spatial aspects of the activity patterns travelers choose (see, e.g., Fig. 1) adds a layer of complexity to the problem of locating charging stations to link the transportation and energy systems. This requires both micro (neighborhood, city, and metropolitan area) and macro (region, state, and nation) driver behavioral dynamics to be studied in detail. If appropriate incentives are developed, drivers could be encouraged not only to act in a way that best serves the grid, but also to act in a way that best serves the transportation system. The incentives could be passive such as pricing electricity for planned contribution at the location of charging facilities (either stations or induction charging embedded in the roadway), or active such as pricing electricity based on congestion in both the power grid and local transportation system. Cognitive and behavioral research is needed to determine the appropriate incentives.

Policy that addresses the planning requirements for charging stations and regulates emerging energy exchange markets is needed. Comprehensive policy that develops joint electricity and transportation programs for incentivizing drivers to participate in the transportation and electricity grid optimization are not yet proposed or even clearly defined.

## Modeling of Complex Systems

To develop policy strategies that allow for faster and more significant penetration of PHEVs/BEVs, research is needed to model the interactive performance of two complex systems, power and transportation, linked through the behavior of individual vehicle operators, where this linkage is determined by the location of interface infrastructure. The behavior of travelers defines the required inputs into power modeling since time-dependent PHEVs/BEVs locations are critical. Every aspect of this meta-system enterprise (power, transport, consumer choice, and infrastructure development) is interlinked, therefore fully understanding policy issues is quite challenging. This section explores each aspect of the modeling approach beginning with transportation modeling, and then power systems modeling, then modeling the role of human agents, and finally determining economic feasibility (see Fig. 2 for illustration).

### Transportation Modeling

Travel models typically contain demand and supply components. While most demand models used in practice are static and consider each leg of a trip separately, activity-based models are gaining momentum. Lemoine et al. [1] illustrate the problems that PHEVs/BEVs could pose if proper incentives are not given to ensure that energy exchange occurs at times beneficial to the grid. Activity-based travel models are better suited for PHEVs/BEVs modeling because they recognize that travel arises from a fundamental need to participate in activities, and thus the models capture trip-chaining behavior (e.g., home to work to grocery to home). Other benefits of activity-based models are the incorporation of intra-household interactions, interpersonal and intrapersonal consistency measures, consideration of space-time constraints on activities and travel, and emphasis on individual level travel patterns (as opposed to monitoring aggregate travel demands). A number of micro-simulation platforms that employ the activity-based paradigm of transportation demand forecasting have been developed in the last 5 years (e.g., [26–28]).

On the supply side, conventional techniques of trip assignment are static in nature, and consider vehicle flows aggregated over one or several hour time periods.

The limitations of the static assignment procedures and the increase in computing capacity have allowed the field to move toward more behaviorally realistic dynamic traffic assignment (DTA) models. DTA techniques offer a number of advantages including capturing the spatial and temporal evolution of traffic dynamics across the transportation network, superior capability to capture traffic congestion buildup and dissipation, and explicitly representing the route-choice effect of external dynamic prices and other costs and incentives. A number of simulation-based DTA modules have been developed in recent years [29–32]. The above mentioned features of DTA make it an ideal choice for modeling the network congestion patterns induced by PHEVs/BEVs usage and their impact on other vehicles.

Travel models produce numerous outputs, metrics, and system properties. Of critical importance for connecting the transportation and energy models are predictions regarding time-dependent vehicle locations. This inference directly relates to the number of PHEVs/BEVs present at a specific power grid node, which will be related to the node's self-admittance described in the next subsection on power systems modeling. Consideration of multiple classes of travelers, PHEVs/BEVs and non-PHEVs/BEVs, will be critical until PHEVs/BEVs reach high percentage penetration.

It has been long understood that through pricing-based incentives, the system-level performance of transportation networks can be greatly improved. The entire field of congestion pricing (e.g., [33, 34]) addresses this fact. For instance, PHEVs/BEVs provide a novel opportunity to achieve gains in controlling and managing congestion in transportation systems through an incentive based approach that persuades users to act in an altruistic manner. Further, such incentives provide a unique opportunity (and complexity) in that dual objectives must be balanced: improving the efficiency of the transportation as well as that of the power system. For the transportation system, incentives influence route, departure, as well as destination choice. Incentives change the fundamental costs traveler's associate with their choices and a new general cost dynamic equilibrium emerges (for normal operating states). This requires a further broadening of the previously mentioned integrated modeling

approach to include generalized costs as well as heterogeneous values of time.

Clearly, there will be significant uncertainty in the model inputs that must be built-in to ensure that the policy recommendations work well for a wide range of potential future outcomes. A vast amount of research has already been performed on stochastic transportation modeling both on the demand and supply side [35–40].

### Power Systems Modeling

The planning, design, and operation of modern power systems call for extensive and detailed simulation. Models used to simulate power system behavior depend on the purpose and uses. When considering the need of studying PHEVs/BEVs impact on power system, different levels of modeling are required.

At the macro level, the power system planning related to the uses of PHEVs/BEVs requires understanding of the generation, storage, and load characteristics, as well as power flow projections impacted by the anticipated use of PHEVs/BEVs. A stochastic nature of PHEV/BEV use in the multiple possible roles will require advanced probabilistic methods for power flow analysis, as well as stochastic optimization related to operation and investment planning of dispersed generation [41, 42]. Enhanced modeling techniques must be developed for PHEV/BEV behavior as a load to assess dynamic stability of the power system operating in G2V mode [43]. Hadley [25] used the Oak Ridge Competitive Electricity Dispatch (ORCED) model to simulate PHEV/BEV electricity demand. It did not directly include transmission and distribution impacts, but discussed the issues of increased continuous transmission. Also, power system contingency analysis must be improved to account for the dynamic nature of both temporal and spatial properties of PHEVs/BEVs. In the V2G mode, PHEVs/BEVs may impact power grid operation in many different roles, both as energy storage used to improve performance of renewable energies such as wind and solar [44], as well as a market participant through aggregated distributed generation [10, 11, 45]. While it has been recognized that PHEVs/BEVs can be used for regulation services [10, 11], some studies also suggested the PHEVs/BEVs use for peak power “shaving” services [46].

A customized modeling tool that allows examining the potential impacts of large scale deployment of PHEVs/BEVs on a given electricity system, such as the “PHEV-load” tool developed by the National Renewable Energy Laboratory (NREL) may be needed [47].

At the micro level, the PHEV/BEV powertrain system itself, which is a very complex dynamic electromechanical system, may be studied. Specialized modeling and simulation tools, such as Argon National Laboratory’s (ANL’s) Powertrain System Analysis Toolkit (PSAT) are well suited for such an analysis [48]. This toolkit allows detailed modeling of charging and discharging dynamics of PHEVs/BEVs, which is crucial when defining properties of PHEVs/BEVs as loads, energy storage, or generation, as discussed above. Other ANL’s tools such as GCtool, GREET, and AirCred may also be needed to assess other impacts [48].

The impact of PHEVs/BEVs ranges from the macro to micro scales, both in size and time. Different power system states (steady state, dynamic, and transient) may need to be represented in a framework using different types of mathematical formulations (waveforms, phasor, and algebraic). This leads to a new requirement for developing a method for linking different modeling techniques for accurate and efficient simulation when representing large scale penetration of PHEVs/BEVs as generators, storage elements, or loads [49].

### Human Behavior Modeling

The widespread adoption of PHEVs/BEVs will place human vehicle operators at the intersection of power and transportation systems. Thus, it is critical to understand human decision making in the context of PHEV/BEV usage and how behavior can be shaped by incentive structures and training interventions. The large disparate group of decision makers includes not only drivers, but also utilities, battery exchange location coordinators, and fleet managers. Cognitive research will be critical to not only to understand and optimize human decision making involving PHEVs/BEVs, but to also increase the rate of PHEV/BEV adoption.

Route planning for any type of vehicle is an example of a dynamic decision task [50]. Choosing a route

requires a series of interrelated decisions that occur in a changing and uncertain environment. PHEVs/BEVs introduce a number of additional decision elements, such as whether to draw energy from the grid or deliver energy to the grid at destinations with facilities allowing such interfaces. Complicating this decision process, G2V costs and V2G credits vary through time and are not perfectly predictable from the driver's perspective.

One successful framework for modeling human performance in dynamic decision making tasks is reinforcement learning (RL) [51, 52]. The theory of RL comprises an array of techniques for learning temporally extended tasks in dynamic environments. An agent is assumed to be immersed in its environment, with some number of actions available to be taken at any given time. The chosen action has an effect, depending on the current state of the environment, the immediate reward (or punishment) the agent receives, and the future state of the environment. Thus actions can influence situations and rewards arbitrarily far into the future, and successful performance hinges on effective planning and coordination of extended sequences of actions [53].

Previous research demonstrates that RL agents and humans are more likely to discover the underlying structure of a task when state cues are present that allow for generalization [54]. A state cue in the context of PHEV/BEV decision making would include observable properties – such as time of day, weather conditions, and congestion – that enable prediction of G2V credits and V2G costs. State cues play a critical role in shaping learning and it has been shown that variability in state predictors disrupts performance more than equivalent variability in the reward structure [55]. Further research is needed to examine how variability in state cues and reward structure affects PHEV/BEV route selection. Establishing how PHEV/BEV driver performance (with respect to improving conditions on the grid and transportation system) declines with variability in state cues is important because transient changes in incentives could have negative, unintended consequences, making it difficult for people to acquire the basic pricing contingencies. Research is also needed to find the best methods for PHEV/BEV operators or aggregators to learn about incentive structures.

Various types of feedback are available (e.g., Reward Only, First Error), and the optimal approach should be determined via experiment.

### Determining Economic Feasibility

To take advantage of the proposed “transportation-energy” markets, interface infrastructure – the facilities that will bridge the two systems and serve as energy transfer points – must be developed and planned. While prevalence of PHEVs/BEVs in the future is unknown, their ultimate value can only come if the interfaces are in place. This leads to a situation where the demand for PHEVs/BEVs depends on the infrastructure supply, which in turn is defined by the demand. The traditional project valuation models fall short of accounting for this feedback loop.

Developing interface infrastructure is a uniquely challenging problem because the equipment must not only adjust energy flow over time, but the location of transfer points must be determined to maximize long-term value and minimize risks. Technology adoption, incentives, and system interdependencies all play a role.

To maximize the value of developing interface infrastructure in a particular location, two aspects of the problem must be considered: (1) the value created to the grid by using PHEVs/BEVs for regulation services, and (2) the value of the activity-based travel patterns that could include a visit to the interface infrastructure. The former value can be explicitly determined, but unless the latter is considered and travelers are enticed to use the new infrastructure, the value to the grid will not be achieved. Typically, the traveling public selects route and activity patterns without considering energy exchange opportunities. New methodologies and modeling techniques must be developed for valuing interface infrastructure given its dependence on traveler behavior.

Unlike most past research into making investment decisions for infrastructure projects that focus on a single system (e.g., [56, 57]), the problem posed here must consider the interdependencies between several systems as well as the rate of technology adoption (availability of PHEVs/BEVs to use this facility and generate value). In fact, this problem exhibits both spatial network effects and strategic “bandwagon”

network externalities (see seminal contributions in this area by Rohlfs [58], Farrell and Saloner [59], and David and Greenstein [60]).

It is clear that in the face of this bandwagon effect, the value of deferral flexibility is marginal. Hence, the project developer action space should consider actions that promote early adoption without fully committing to irreversible capital expenditures. Stochastic modeling approaches could be useful here to consider that the outcome and uncertainty space of the valuation problem is decision dependent (see, e.g., [61]).

### Benefits

This section aims at demonstrating the potential benefits of PHEVs/BEVs that may be used to feed power back to home or office building, which is known as “Vehicle-to-Building” (V2B) operation. The new parking facility called “smart garage” is introduced and its eclectic power capacity is discussed. Based on the availability analysis of smart garages, a strategy to adopting PHEV/BEV uses in the V2B mode under peak load and outage condition is proposed. V2B approach considers PHEVs/BEVs as a generation resource for the buildings at certain periods of time via bidirectional power transfers, which could increase the flexibility of the electrical distribution system operation. It is expected that V2B operation will improve the reliability of the distribution system, provide extra economic benefits to the vehicle owners, and reduce the home or building electricity purchase cost based on the demand side management and outage management programs with customer incentives.

### Demand Side Management (DSM)

For electric utility, DSM is defined as “Utility-sponsored programs to influence the time of use and amount of energy use by select customers,” which includes peak clipping, valley filling, load shifting, strategic conservation, strategic load growth, and flexible load shape [62]. However, for utility end-user (customer), DSM is often understood to include two components: energy efficiency (EE) and demand response (DR). EE is designed to reduce electricity consumption during all hours of the year; DR is designed to change

**PHEVs and BEVs in Coupled Power and Transportation Systems. Table 1** DSM benefits to customer, utility, and society [64]

Customer benefits	Societal benefits	Utility benefits
Satisfy electricity demands	Reduce environmental impacts	Lower cost of service
Reduce/stabilize costs	Conserve resources	Improved operating efficiency
Improve value of service	Protect global environment	Flexibility of operation
Maintain/improve lifestyle	Maximize customer welfare	Reduced capital needs

on-site demand for energy in intervals and associated timing of electric demand by transmitting changes in prices, load control signals, or other incentives to end users to reflect existing production and delivery costs [63]. By cooperative activities between the utility and its customers to utilize DSM, it will provide the benefits to the customer, utility, and society as a whole, which is summarized in Table 1 [64].

In the V2B option, the owners will plug in their vehicles during the day at their final destination for a given time frame. As an example, this may be either at their workplace (central business district) or at the place of their study (university). The destinations, either parking lots or parking garages next to the buildings, are assumed to be equipped with a bidirectional charger and controller. The parking facility should allow either charge or discharge mode for the car batteries when necessary. The idea is that the parking facility can offer an aggregation service for charging the batteries when the building demand is lower than its peak load and discharge the batteries to partially supply the building to reduce the peak demand during a high demand. This mode will be considered as DSM by V2B. Considering the electricity rate when the vehicle batteries were charged is lower than when the batteries are discharged, the battery storage may be used to offset high cost during the peak demand.

### Outage Management (OM)

Another important benefit of V2B is using the battery energy storage in PHEVs/BEVs as an emergency backup power for the commercial facility/building, which increases the reliability of the power supply for that load.

An outage is typically caused by several unplanned events, and a timely detection and mitigation of such situations is a real concern for the utility. Outage management system helps the operators to locate an outage, repair the damage, and restore the service. Outage management must be performed very quickly to reduce outage time. Recently completed project proposes an optimal fault location scheme which will help the operator to find the faulted section very quickly [65]. In this entry, the restoration strategy under an outage will be mainly discussed.

The following types of outages and studies about the impact of PHEVs/BEVs adoption are considered:

- (a) *Outage beyond the distribution system*: These may be caused by generator failure, fault in transmission line, or substation busbar. Usually spinning reserves are kept for these circumstances. From the previous studies it is concluded that PHEVs/BEVs can be a candidate solution for spinning reserves (as the traditional fastest acting spinning reserve generators are highly costly while PHEVs/BEVs qualify for fast response with lesser cost). One may consider using a real-time security constrained optimal power flow under the contingencies to calculate the amount of PHEV/BEV battery capacity required for a certain location at a specific instance.
- (b) *Outage in distribution system*: These may be caused by fault inside the distribution system and can be mitigated by precise spatial adjustment of PHEV/BEV battery generation that may be used to feed electricity locally during and after outage.

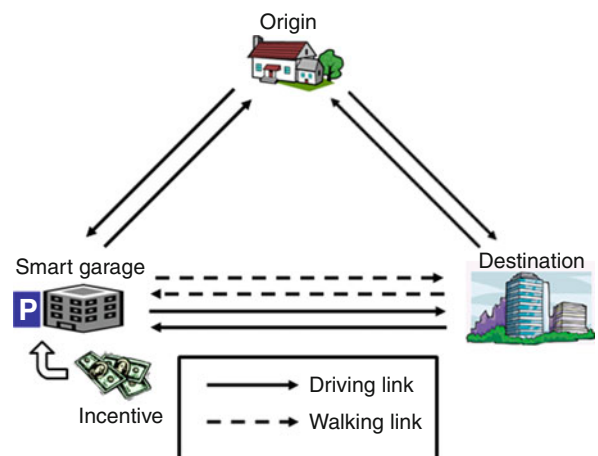
To propose the restoration strategy where PHEV/BEV batteries are used to mitigate an outage condition, the information about events (where the fault is located and how the impact will propagate) and the location of the battery storage need to be correlated. Thus, a spatial as well as temporal analysis should be performed.

The restoration strategy can be executed in the following steps:

1. Detect a fault.
2. Estimate the location of the fault.
3. Analyze the amount of battery generation required and the availability of PHEVs/BEVs that can provide an alternative generation support in the vicinity of the faulted area until the faulted section is repaired. This will also consider the generation duration requirement (i.e., time to repair the faulted section).
4. Schedule the aggregated PHEVs/BEVs generation optimally. This is a multi-objective optimization problem which can be formulated to minimize the distance between location of the fault and available PHEVs/BEVs battery generation locations as well as minimize the operating cost under system operation and security constraints.

### Garage Location and Charging/Discharging Infrastructure

Commercial and public parking garages in a central business district (CBD) provide thousands of parking spaces for commuters and visitors. After penetrating the conventional vehicle market, owners of PHEVs/BEVs will be using these parking garages, which may



PHEVs and BEVs in Coupled Power and Transportation Systems. Figure 6

Simple transportation network with smart garage



provide an aggregated service to act as an electric power source or storage.

Figure 6 shows a simple transportation network with smart garage building. As a smart garage is constructed, PHEV/BEV drivers have two options: proceed to final destination directly or park at the smart garage and walk to the destination along walking links. Drivers in transportation network select parking garage based on the location and financial incentives (less parking fee), which can be modeled as traffic assignment problem. Demand of smart garage (number of parked PHEVs/BEVs) calculated from the traffic assignment problem would vary by the location and incentive of the smart garage.

Electric power capacity of smart garage is estimated based on demand of smart garage. Demand of smart garage building is not constant. Generally, the demand of smart garage building during the day would be higher than during the night, similar to the demand structure for a conventional garage as shown in Fig. 7. Due to the versatility, electric power capacity needs to be defined in two parts: for periodic service and for continuous service as in Fig. 7. The available electric power estimated based on the demand of smart garage can be used for determining the support service that can be provided during outage management and demand side management in vehicle-to-building (V2B) mode.

## Case Study

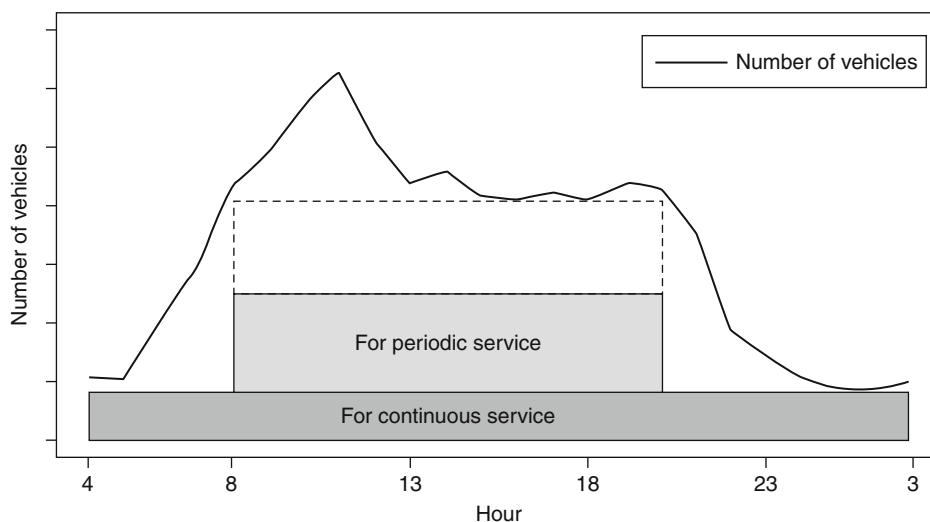
Test cases for two scenarios are studied: demand side management using V2B mode during peak power demand and outage management using V2B mode during faults.

### Demand Side Management During Peak Power Demand

In this case, a large commercial building is analyzed to demonstrate the potential savings using demand side management based on V2B operation. Itron, Inc. prepared a technical survey for the California Energy Commission (CEC), which modeled difference commercial sectors, including large office building [66]. The load shapes include typical day, hot day, cold day, and weekend for each of four seasons. According to the definition used in this report, large office buildings are defined as premises with total floor area equal or larger than 30,000 sq ft. The largest electric end-uses in this building type are interior lighting, cooling, office equipment, and ventilation [66].

The summer typical load shape for a large office building is selected for our case study. The single building demand is obtained from the results reported in the literature [66]. The following assumptions are taken:

- The studied building is 450,000 sq ft.
- There are up to 80 PHEVs/BEVs that arrive at 8 a.m. and are available for the entire day.

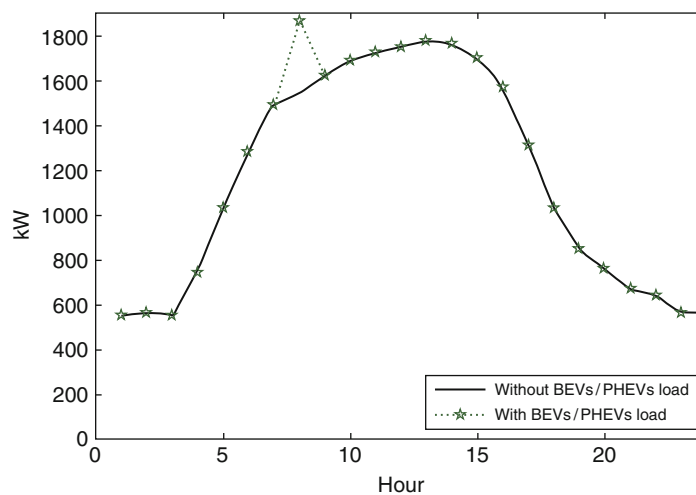


PHEVs and BEVs in Coupled Power and Transportation Systems. Figure 7  
Demand of smart garage for a day

- Maximum capacity of each vehicle is 10 kWh (very conservative for BEVs).
- The batteries in PHEVs/BEVs are drained by an average of 4.0 kWh by the driving cycle used.

When PHEVs/BEVs are on site, the building can charge the batteries during the morning hours (lower electricity price) and drain the batteries by an equal amount during afternoon hours (higher electricity price). Thus the owner of PHEV/BEV will have the required energy in his/her battery to make sure the driving cycle to return home can be met. Figure 8 shows the impacts of charging PHEVs by faster charging methods (AC Level 3 or DC charging). It will elevate the peak demand to 1.86 MW of the office building since the faster charging method will cause a large load in a short period (10–15 min), which is not recommend for either utilities or customers.

Figure 9 shows the change in the load shape for the typical summer day by using the AC Level 1 charging method defined by the Society for Automotive Engineers (SAE) J1772 [67]. The load curve was changed by shifting the afternoon peak load to the morning off-peak load when charging and discharging PHEVs/BEVs. Considering the rate structures for peak and off-peak load in commercial buildings, peak load shifting using V2B mode may provide the electricity bill saving. Further study could be conducted to show the total saving expressed in dollars.



**PHEVs and BEVs in Coupled Power and Transportation Systems. Figure 8**  
Impacts of faster charging PHEVs/BEVs on load demand

**Outage Management During Faults** The proposed restoration scheme was tested on a small distribution system (IEEE 37 node radial test feeder [68]).

This is an actual feeder located in California, which consists of several unbalanced spot loads. The nominal voltage is 4.8 kV.

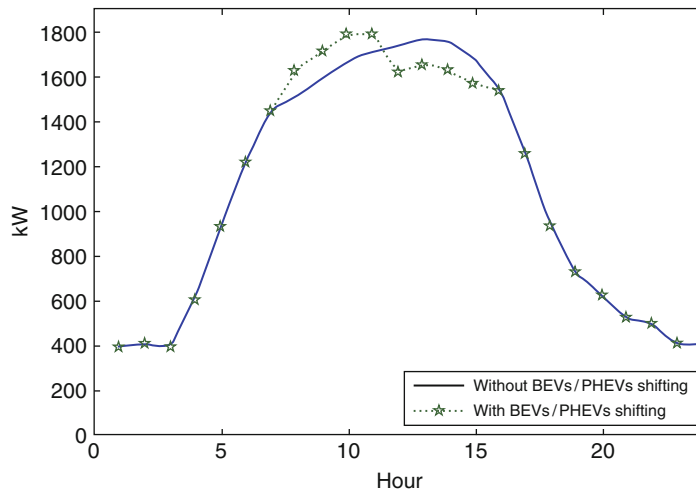
Figure 10 shows the test feeder with smart garages at some nodes.

The following assumptions are taken:

- Three nodes are specified as smart garages (nodes 718, 735, and 740).
- Maximum capacity of each vehicle is 10 kWh.
- Discharge vehicles with state of charge (soc >70%).
- PHEV/BEV tariff for charging is 5 c/kWh and for discharging is (15–40) c/kWh (depending on different garages). Discharging tariff for node 718 is 40 c/kWh, for node 735 is 30 c/kWh, and for node 740 is 25 c/kWh.

Under normal operating condition, node no. 799 acts as an infinite bus and all the loads are fed through it. Two different outage cases are studied:

1. Case 1: Fault on or beyond node 799: PHEVs/BEVs at nodes 718, 735, and 740 were scheduled to satisfy all the loads on the feeder. Table 2 shows the case results.
2. Case 2: Fault on line segment 703–730: Node 799 will supply all the loads beyond this line segment.



**PHEVs and BEVs in Coupled Power and Transportation Systems. Figure 9**  
Peak load shifting with PHEVs/BEVs for a typical summer daily load

PHEVs/BEVs at nodes 735 and 740 will be scheduled to satisfy the island created by a fault on line 703–730. Table 3 shows the case results.

### Conclusions and Future Research

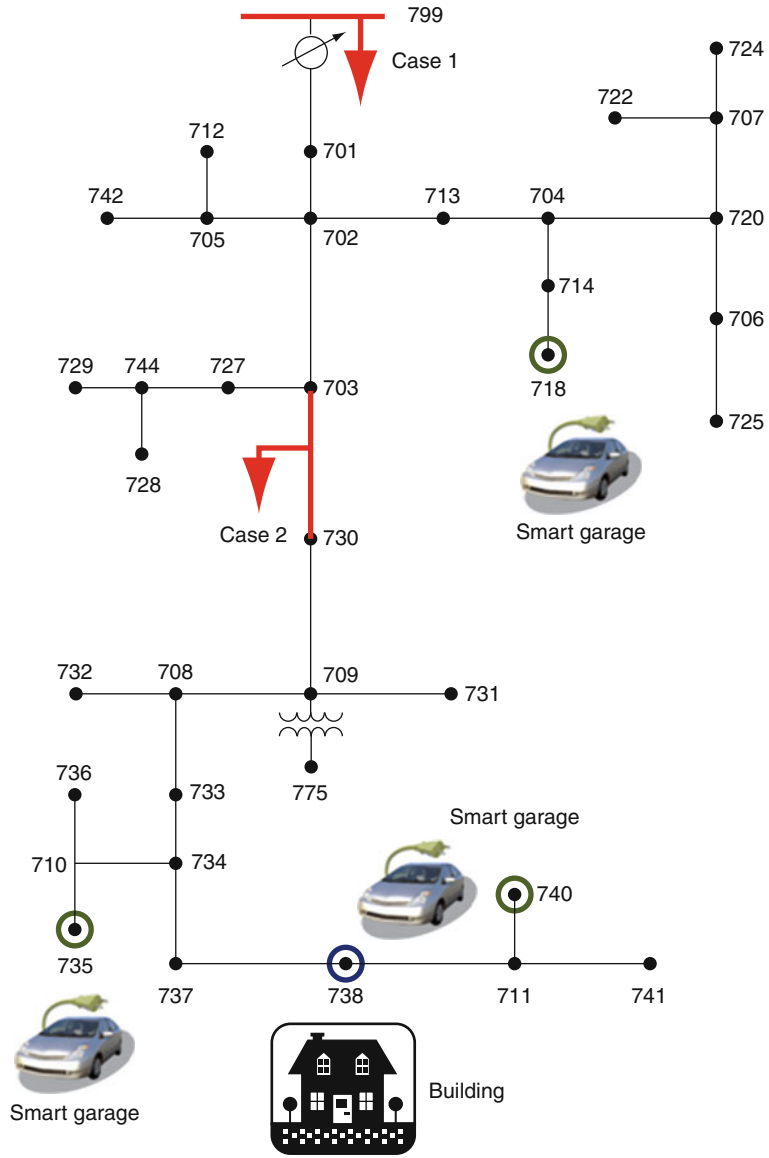
The policy implications of widespread PHEV/BEV deployment in the energy and transportation systems are explored. Previous research has approached the problem from selected angles, making many simplifying assumptions. Some thoughts on how the problem may be approached from a non-myopic perspective are provided.

In summary, numerous policy shifts are needed to realize the full potential of PHEVs/BEVs, and the cooperation of the transportation and energy sectors is vital. If policies such as the ones outlined in this paper are adopted, PHEVs/BEVs can provide many benefits to the electric grid in terms of reliability and stability by acting as mobile decentralized storage and allowing for vehicle-to-grid and grid-to-vehicle services. PHEVs/BEVs will also allow for enhanced penetration of renewable energy resources such as wind and solar, which will also aid with energy security by reducing dependence on foreign sources of oil. Important benefits can be made to air quality through transferring pollution from numerous mobile sources to fewer point sources that are easier to control and may

participate in cap-and-trade markets. In addition to the carbon market, new markets will be created in power systems due to the potential for PHEVs/BEVs (or aggregators of PHEVs/BEVs) to participate, particularly with ancillary and regulation services. Lastly, charging stations must be planned and developed carefully to allow for flexibility in driver options and optimal performance of the transportation and electricity networks.

The proposed multi-layered modeling framework considers the spatial and temporal nature of the system interactions. PHEV/BEV time-dependent travel patterns are outputs of a transportation model and inputs to power systems model. The framework also includes cognitive behavior modeling for the purposes of developing appropriate incentives to encourage drivers to behave in a way that improves the efficiency of the transportation and energy systems.

The potential benefits of using PHEVs/BEVs as dynamically configurable dispersed energy storage that can serve as load or generation in a power system as needed is discussed. If serving in G2V as well as V2B mode and if aggregated, PHEVs/BEVs may play a major role in both the electricity and the transportation networks. Selecting garage location and charging/discharging infrastructure needs special attention from the transportation system demand point of view. For demand side management in electricity networks, the



PHEVs and BEVs in Coupled Power and Transportation Systems. Fig. 10  
Diagram of test feeder with smart garages

PHEVs and BEVs in Coupled Power and Transportation Systems. Table 2 Case study 1: results for PHEV/BEV generation scheduling

Node 718			Node 735			Node 740		
Ph-1 (kW)	Ph-2 (kW)	Ph-3 (kW)	Ph-1 (kW)	Ph-2 (kW)	Ph-3 (kW)	Ph-1 (kW)	Ph-2 (kW)	Ph-3 (kW)
0	0	411	300	300	300	427	339	380

Total cost for three phases is \$733.2/h

**PHEVs and BEVs in Coupled Power and Transportation Systems. Table 3** Case study 2: results for PHEV/BEV generation scheduling

Node 735			Node 740		
Ph-1 (kW)	Ph-2 (kW)	Ph-3 (kW)	Ph-1 (kW)	Ph-2 (kW)	Ph-3 (kW)
300	127	300	51	0	81

Total cost for three phases is \$221.35/h

use of PHEVs/BEVs to create a peak load shifting strategy can reduce the electricity purchase cost for the customer and vehicle owner. For outage management in electricity networks, the use of PHEVs/BEVs to generate power during outage restoration stage is envisioned by solving a multi-objective optimization problem of merit-order scheduling of PHEVs/BEVs under operating constraints.

In recent years, Smart Grid revolution has begun with the sponsorship and involvement from government, businesses, utilities, and other stakeholders, especially with the development and integration of renewable energy resources. Envisioning the longer-term impact, if there is enough aggregated PHEV/BEV vehicles, such as a fleet, they can serve as backup generation and storage for renewable energy in smart grid applications. Many other functions of the future electricity network may be affected when PHEVs/BEVs act as dynamically configurable energy storage, which may have profound impact on the transportation networks as well. Better understanding of the role of PHEVs/BEVs in coupled power and transportation systems will be beneficial to transform existing power grid into the Smart Grid, a power system that is more efficient, reliable, resilient, and responsive.

### Acknowledgments

The author wishes to acknowledge numerous colleagues and graduate students who contributed to the findings of this entry: Dr. Bradley Love, and Dr. Jennifer Duthie from The University of Texas at Austin, as well as Dr. Ivan Damnjanovic, and graduate students Mr. Chengzong Pang, Ms. Papiya Dutta, and Mr. Seok Kim from Texas A&M University. Funding for this study came from the National Science Foundation

through I/UCRC grant for the Center for “PHEVs/BEVs: Transportation and Electricity Convergence,” and another NSF I/UCRC grant for the “Power Systems Engineering Research Center.”

### Bibliography

#### Primary Literature

1. Lemoine DM, Kammen DM, Farrell AE (2008) An innovation and policy agenda for commercially competitive plug-in hybrid electric vehicles. *Environ Res Lett* 3(1):014003
2. Schrank D, Lomax T (2007) The 2007 urban mobility report. Texas Transportation Institute
3. U.S.-Canada Power System Outage Task Force (2004) Final report on the August 14, 2003 Blackout in the United States and Canada: causes and recommendations. <http://www.nerc.com>
4. Burges K, Twele J (2005) Power systems operation with high penetration of renewable energy – the German case. In: 2005 international conference on future power systems, Amsterdam, pp 1–5
5. Kelso JAS (1995) Dynamic patterns: the self-organization of brain and behavior. MIT Press, Cambridge, MA
6. Turvey MT, Moreno M (2006) Physical metaphors for the mental lexicon. *Ment Lexicon* 1:7–33
7. Van Orden GC, Holden JG, Turvey MT (2003) Self-organization of cognitive performance. *J Expr Psychol Gen* 132:331–350
8. Brooks AN (2002) Vehicle-to-grid demonstration project: grid regulation ancillary service with a battery electric vehicle. AC Propulsion, California
9. Duncan R, Osborne MJ (2005) Report on transportation convergence, Austin energy
10. Kempton W, Tomic J, Letendre S, Brooks A, Lipman T (2001) Vehicle-to-grid power: battery, hybrid, and fuel cell vehicles as resources for distributed electric power in California. Report # UCD-ITS-RR-01-03, Electric transportation program
11. Kempton W, Tomic J (2005) Vehicle to grid implementation: from stabilizing the grid to supporting large-scale renewable energy. *Power Sources* 144(1):280–294
12. Kintner-Meyer M, Schneider K, Pratt R (2007) Impacts assessment of plug-in hybrid vehicles on electric utilities and regional U.S. power grids, part 1: technical analysis
13. Solomon J, Vincent R (2003) Development and evaluation of a plug-in HEV with vehicle-to-grid power flow. Final report, A.C. Propulsion
14. Letendre S, Perez R, Herig C (2002) Battery-powered, electric-drive vehicles providing buffer storage for PV capacity value. In: Proceedings of the 2002 American solar energy society annual conference, Boulder
15. Kempton W, Tomic J (2005) Vehicle-to-grid power fundamentals: calculating capacity and net revenue. *J Power Source* 144(1):268–279

16. Kempton W, Tomić J (2005) Vehicle-to-grid implementation: from stabilizing the grid to supporting large-scale renewable energy. *J Power Source* 144(1):280–294
17. Hadley SW, Tsvetkova A (2008) Potential impacts of plug-in hybrid electric vehicles on regional power generation. Oak Ridge National Laboratory, Oak Ridge, TN, ORNL/TM-2007/150. [http://apps.ornl.gov/~pts/prod/pubs/ldoc7922\\_regional\\_phev\\_analysis.pdf](http://apps.ornl.gov/~pts/prod/pubs/ldoc7922_regional_phev_analysis.pdf)
18. Meliopoulos S, Meisel J, Cokkinides G, Overbye T (2009) Power system level impacts of plug-in hybrid vehicles, Pserc project T34 final report #09-12. [http://www.pserc.wisc.edu/documents/publications/reports/2009\\_reports/](http://www.pserc.wisc.edu/documents/publications/reports/2009_reports/)
19. Andersson SL, Elofsson AK, Galus MD, Goransson L, Karlsson S, Johnsson F, Andersson G (2010) Plug-in hybrid electric vehicles as regulating power providers: case studies of Sweden and Germany. *Energy Policy* 38(6):2751–2762
20. Guille C, Gross G (2009) A conceptual framework for the vehicle-to-grid (V2G) implementation. *Energy Policy* 37(11):4379–4390
21. Lerner EJ (2003) What's wrong with the electric grid? *The industrial Physicist*, Oct/Nov 2003
22. Wardrop JG (1952) Some theoretical aspects of road traffic research. *Proc Inst Civil Eng* 1(2):325–378
23. U.S. Department of Transportation, Federal Highway Administration, Office of Highway Policy Information (2007) Traffic volume trends. <http://www.fhwa.dot.gov>
24. Toole-Holt L, Polzin SE, Pendyala RM (2005) Two minutes per person per day each year: exploration of growth in travel time expenditures. *Transp Res Rec* 1917:45–53
25. Hadley SW (2006) Impact of plug-in hybrid vehicles on the electric grid. Oak Ridge National Laboratory, for the U.S. department of energy, ORNL/TM-2006/554
26. Bhat CR, Guo JY, Srinivasan S, Sivakumar A (2004) A comprehensive econometric microsimulator for daily activity-travel patterns. *Transp Res Rec* 1894:57–66
27. Vovsha P, Petersen E, Donnelly R (2004) Impact of intra-household interactions on individual daily activity-travel patterns. In: Proceedings of the 83rd annual meeting of the transportation research board, Washington, DC
28. Bowman JL, Bradley MA (2005) Activity-based travel forecasting model for SACOG, technical memos numbers 11–11. <http://jbowman.net>
29. Waller ST, Ziliaskopoulos AK (1998) A visual interactive system for transportation algorithms. In: Proceedings of the 78th annual meeting of the transportation research board, Washington, DC
30. Taylor NB (1990) CONTRAM 5: an enhanced traffic assignment model. TRLL research report 249
31. Ben Akiva M, Bierlaire M, Koutsopoulos H, Mishalini R, DynaMIT (1998) A simulation-based system for traffic prediction and guidance generation. Presented at TRISTAN III, Delft, The Netherlands
32. Mahmassani HS, Hu TY, Jayakrishnan R (1992) Dynamic traffic assignment and simulation for advanced network informatics (DYNASMART). In: Proceedings of the second international capri seminar on urban traffic networks, Italy
33. Arnott R, Small K (1994) The economics of traffic congestion. *Am Scientists* 20:123–127
34. Hearn DW, Ramana MV (1998) Solving congestion toll pricing models. In: Marcotte P, Nguyen S (eds) *Equilibrium and advanced transportation modeling*. Kluwer, Boston, pp 109–124
35. Morrison SA (1986) A survey of road pricing. *Transp Res* 20A(2):87–97
36. Waller ST, Schofer JL, Ziliaskopoulos AK (2001) Evaluation with traffic assignment under demand uncertainty. *Transp Res Rec* 1771:69–75
37. Karoonsoontawong A, Waller ST (2006) Dynamic continuous network design problem: linear bi-level programming and metaheuristic approaches. *Transp Res Rec* 1964:104–117
38. Ukkusuri S, Tom VM, Waller ST (2007) Robust network design problem under demand uncertainty. *Comput-Aided Civ Infrastruct Eng* 22:6–18
39. Waller ST, Ziliaskopoulos AK (2006) A chance-constrained based stochastic dynamic traffic assignment model: analysis, formulation and solution algorithms. *Transp Res C: Emerg Technol* 14(6):418–427
40. Duthie J, Unnikrishnan A, Waller ST (2006) Network evaluation with uncertain and correlated long-term demand. In: Proceedings of 85th transportation research board meeting, Washington, DC
41. Stefopoulos G, Meliopoulos APS, Cokkinides G (2005) Probabilistic power flow with non-conforming electric loads. *Int J Electr Power Energy Syst* 27(9–10):627–634
42. Gollmer R, Neise U, Schultz R (2007) Risk modeling via stochastic dominance in power systems with dispersed generation. Technical report, Department of mathematics, University of Duisburg, Germany
43. Choi BK, Chiang HD, Li YH (2006) Measurement-based dynamic load models: derivation, comparison, and validation. *IEEE Trans Power Syst* 21:1689–1697
44. Short W, Denholm P (2006) A preliminary assessment of plug-in hybrid electric vehicles on wind generation markets. Technical report, NREL/TP-620-39729
45. Tomic J, Kempton W (2007) Using fleets of electric drive vehicles for grid support. *J Power Sources* 168(2):459–468
46. Kempton W, Kubo T (2000) Electric-drive vehicles for peak power in Japan. *Energy Policy* 28:9–18
47. Denholm P, Short W (2006) An evaluation of utility system impacts and benefits of optimally dispatched plug-in hybrid electric vehicles. Technical report, NREL/TP-620-40293
48. Argon National Laboratory Tools. <http://www.transportation.anl.gov/software/PSAT/index.html>
49. Kasztenny B, Kezunovic M (2000) A method for linking different modeling techniques for accurate and efficient simulation. *IEEE Trans Power Syst* 15(1):65–72

50. Busemeyer JR (2002) Dynamic decision making. In: Smelser NJ, Baltes PB (eds) *International encyclopedia of the social and behavioral sciences*, vol 6. Elsevier, Oxford, pp 3903–3908
51. Fu WT, Anderson JR (2006) From recurrent choice to skill learning: a reinforcement-learning model. *J Exp Psychol Gen* 135:184–206
52. Gray WD, Sims CR, Fu W-T, Schoelles MJ (2006) The soft constraints hypothesis: a rational analysis approach to resource allocation for interactive behavior. *Psychol Rev* 113:461–482
53. Sutton RS, Barto AG (1998) *Reinforcement learning: an introduction*. MIT Press, Cambridge, MA
54. Gureckis TM, Love BC (2009) Short term gains, long term pains: reinforcement learning in dynamic environments. *Cognition* 113:293–313
55. Gureckis TM, Love BC (2009) Learning in noise: dynamic decision-making in an uncertain environment. *J Math Psychol* 53:180–193
56. Zhao T, Fu CC (2006) Infrastructure development and expansion under uncertainty: a risk-preference-based lattice approach. *ASCE J Constr Eng Manag* 132(6):620–625
57. Zhang Z, Damjanovic I (2006) Quantification of risk cost associated with short-term warranty-based specifications for pavements. *Transp Res Rec* 1946:3–11
58. Rohlfs JH (1974) A theory of interdependent demand for a communication service. *Bell J Econ Manag Sci* 5(1):16–74
59. Farrell J, Saloner G (1985) Standardization, compatibility and innovation. *Rand J Econ* 16:70–82
60. David PA, Greenstein S (1990) The economics of compatibility standards: an introduction to recent research. *Econ Innovat New Tech* 1:3–41
61. Jonsbraten TW, Wets RJB, Woodruff DL (1998) A class of stochastic programs with decision dependent random elements. *Ann Oper Res* 82:83–106
62. Gellings CW (1985) The concept of demand-side management for electric utilities. *Proc IEEE* 73(10):1468–1470
63. NERC (2007) Data collection for demand-side management for quantifying its influence on reliability: results and recommendations. In: North American electric Reliability Corporation, Princeton. [http://www.nerc.com/docs/pc/drdrtf/NERC\\_DSMTF\\_Report\\_040308.pdf](http://www.nerc.com/docs/pc/drdrtf/NERC_DSMTF_Report_040308.pdf)
64. IIEC (2006) Demand side management best practices guidebook or pacific island power utilities. International Institute for Energy Conservation, Washington, DC. [www.sidsnet.org/docshare/other/20070110DSMBestpractices.pdf](http://www.sidsnet.org/docshare/other/20070110DSMBestpractices.pdf)
65. Kezunovic M, Ward J et al (2009) Integration of asset and outage management tasks for distribution systems. Pserc project T36 final report #09-11. [http://www.pserc.wisc.edu/documents/publications/reports/2009\\_reports/](http://www.pserc.wisc.edu/documents/publications/reports/2009_reports/)
66. Itron, Inc. (2006) California commercial end-use survey: consultant report. CEC-400-2006-005, California Energy Commission. <http://www.energy.ca.gov/2006publications/CEC-400-2006-005/CEC-400-2006-005.PDF>
67. SAE (2010) Recommended practice for electric vehicle and plug in hybrid electric vehicle conductive charger coupler, SAE standard J1772, Jan 2010
68. Radial Test Feeders – IEEE distribution system analysis sub-committee. <http://ewh.ieee.org/soc/pes/dsacom/testfeeders.html>

## Books and Reviews

- Dowds J, Hines P et al (2010) Plug-in hybrid electric vehicle research project: phase two report. Burlington, VT, Transportation research center. UVM TRC report #:10-001. <http://www.uvm.edu/~trans...reports/UVM-TRC-10-001.pdf>
- Galus MD, Andersson G (2008) Demand management of grid connected plug-in hybrid electric vehicles (PHEV). In: Energy 2030 conference, ENERGY 2008. IEEE, 17–18 Nov 2008, pp 1–8
- Galus MD, Andersson G (2009) Integration of plug-in hybrid electric vehicles into energy networks. PowerTech, 2009 IEEE bucharest. June 28–July 2 2009, pp 1–8
- Galus MD, Andersson G (2009) Power system considerations of plug-in hybrid electric vehicles based on a multi energy carrier model. In: Power & energy society general meeting, PES '09. IEEE. 26–30 July 2009, pp 1–8
- Kempton W, Udo V et al (2008) A test of vehicle-to-grid (V2G) for energy storage and frequency regulation in the PJM system. University of Delaware, Pepco Holdings, Inc, PJM interconnect, and Green Mountain College. [http://www.magicconsortium.org/\\_Media/test-v2g-in-pjm-jan09.pdf](http://www.magicconsortium.org/_Media/test-v2g-in-pjm-jan09.pdf)
- Markel T, Simpson A (2005) Energy storage systems considerations for grid-charged hybrid electric vehicles. In: Vehicle power and propulsion, 2005 IEEE conference. 7–9 Sept 2005. <http://www.nrel.gov/vehiclesandfuels/vsa/pdfs/38538.pdf>
- Morrow K, Karner D et al (2008) Plug-in hybrid electric vehicle charging infrastructure review, Idaho National Laboratory. Report #: INL/EXT-08-15058. <http://avt.inel.gov/pdf/phev/pevInfrastructureReport08.pdf>
- ORNL (2010) Plug-in hybrid value proposition study final report, Oak Ridge National Laboratory (ORNL). Report #: ORNL/TM-2010/46. <http://www.sentech.org/phev/pdfs/PHEV%20Value%20Proposition%20Study%20Final%20Report%20Draft.pdf>
- Sullivan JL, Salmeen IT et al (2009) PHEV marketplace penetration: an agent based simulation. University of Michigan, Ann Arbor, Transportation Research Institute. Report #: UMTRI-2009-32. <http://deepblue.lib.umich.edu/bitstream/2027.42/63507/1/102307.pdf>
- Turton H, Moura F (2008) Vehicle-to-grid systems for sustainable development: an integrated energy analysis. *Technol Forecasting Soc Change* 75(8):1091–1108
- Wynne J (2009) Impact of plug-in hybrid electric vehicles on California's electricity grid. Masters of environmental management degree, Duke University

## Phosphoric Acid Fuel Cells for Stationary Applications

SRIDHAR V. KANURI, SATHYA MOTUPALLY  
UTC Power, South Windsor, CT, USA

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Market Requirements  
Phosphoric Acid Fuel Cells  
Cell Stack Assembly Components  
Cell Stack Assembly Life  
PAFC Applications  
Future Directions  
Acknowledgments  
Bibliography

### Glossary

**ADG** Anaerobic digester gas.  
**BOP** Balance of plant. Involves components other than fuel cell stacks in a power plant.  
**Bubble pressure** Ability of a component filled with acid to withstand a given pressure of gas.  
**Carbonization** This process involves heating resin-impregnated material to  $\sim 1,000^\circ\text{C}$  to carbonize.  
**CH<sub>4</sub>** Methane.  
**CHP** Combined heat and power. Equipment that generates both electrical and thermal energy.  
**Cloud tower** Equipment to deposit catalyst onto GDL.  
**ECA** Electrochemical area, ideally the Pt surface area available for oxygen reduction or Hydrogen oxidation reaction.  
**Efficiency** Energy output/Energy input.  
**ETU** Electrolyte take-up: Quantity of electrolyte ( $\text{H}_3\text{PO}_4$ ) taken up by a unit weight of carbon.  
**FEP** Fluorinated ethylene propylene.  
**Floc** Mixture of carbon-coated catalyst and PTFE<sup>®</sup>.  
**GDL** Gas diffusion layers.  
**GDL** Gas diffusion layers or substrates.  
**Graphitization** This process involves in heating carbon material to temperatures of  $2,500\text{--}3,000^\circ\text{C}$  to improve thermal conductivity and corrosion resistance.

**H<sub>2</sub>** Hydrogen.  
**H<sub>3</sub>PO<sub>4</sub>** Phosphoric acid.  
**HT-PEM** High-temperature polymer electrolyte membrane fuel cell.  
**Ionic resistance** Resistance for the flow of  $\text{H}^+$  through the electrolyte matrix.  
**kW** Kilo watts.  
**NG** Natural gas.  
**O<sub>2</sub>** Oxygen.  
**PAFC** Phosphoric acid fuel cell.  
**PAN** Polyacrylonitrile.  
**Performance decay** Loss of fuel cell performance due to kinetic, ionic, or mass transport losses.  
**PTFE<sup>®</sup>** Polytetrafluoroethylene.  
**SiC** Silicon Carbide.

### Definition of the Subject

Fuel cells generate power by electrochemically combining fuel such as hydrogen and oxidant such as oxygen in air to produce electrical and thermal energy. Fuel cells generally consist of an anode electrode where fuel is oxidized and cathode electrode where oxygen in air is reduced. The electrolyte which is usually placed between the two electrodes acts as a medium to transport charge carriers (e.g.,  $\text{H}^+$ ,  $\text{CO}^-$ ). Fuel cells are particularly interesting as energy generating devices because they consume reactants without combustion, thus providing higher efficiencies and avoiding the issue of pollution. A fuel cell reaction typically produces water as a by-product which is usually removed from the cell by reactant exhaust.

There are various types of fuel cells that are under development. The most noticeable ones are polymer electrolyte membrane (PEM) fuel cells, phosphoric acid fuel cells (PAFC), molten carbonate fuel cells (MCFC), and solid oxide fuel cells (SOFC). PEM fuel cells are mainly being targeted toward transportation needs due to their ability to provide high power densities at reasonable operating temperatures ( $\sim 100^\circ\text{C}$ ). PAFCs and MCFCs are being developed primarily for stationary applications since their power densities are lower than PEM. SOFCs are currently being developed for both stationary applications and transportation applications but high-temperature material development is needed before they become commercially viable.



## Introduction

Medium-temperature fuel cells can be classified as those operating between 120°C and 250°C. Two main categories of fuel cells fall in this category. One is phosphoric acid fuel cell and the other is high-temperature polymer electrolyte membrane fuel cells (HT-PEMFC). While PAFC is the only fuel cell technology that has demonstrated over 70,000 h of field operation it has cost challenges that have to be overcome for commercialization. On the other hand, significant amount of government and private funding is being devoted to the development of HT-PEMFC for automotive and stationary applications [47]. So far HT-PEM has shown little progress in lab-scale evaluation and has a long road to full-scale demonstration.

This temperature range is particularly interesting because it allows the fuel cell to run on reformed fuel and have tolerance to CO which is the main fuel cell poison generated in the reforming reaction. Most importantly, these temperatures allow for combined heat and power (CHP) capability that can be used in most commercial buildings for either space heating, hot water generation, cooling applications, etc. Today, commercial CHP is the best application for fuel cells in the stationary space due to price of electricity, price of gas, the opportunity for customer energy savings, and the impact of such savings on world energy consumption. Fuel cells in CHP applications need to be efficient (80–90% efficiency), cost effective (less than \$2,000/kW installed), have useable heat (temperatures in the neighborhood of 150–250°C), and long life times (10 years) to provide the required value proposition to customers. PAFCs fit very well with the above requirements except for cost which is being addressed currently by both industry and government.

## Market Requirements

Stationary fuel cell applications include commercial buildings such as supermarkets, data centers, schools, hotels, and hospitals, and industrial users such as chemical plants and refineries and distribution utilities. The primary driver for acceptance of any energy application is the payback period associated. For reference purposes, the payback period in the energy industry is usually on the 3–5 year time scale. In addition to

payback period, strategic factors such as grid congestion and unreliability, rising energy costs, urbanization, global warming, ability to use waste heat, and avoidance of peak load constraints are also factoring more and more into the decision making of customers evaluating fuel cells for stationary applications.

Payback period is usually a function of initial cost of the system and the life cycle costs associated with the system. Initial cost of a PAFC system involves the cost of fuel cell stacks and balance of plant components, cost of integration and assembly of these components and factory acceptance test. Life cycle costs are primarily a function of the efficiency of the system, cost of fuel and maintenance costs. Efficiency in the case of CHP applications involves both electrical and thermal efficiency.

PAFCs have a payback period of 3–5 years (with various government incentives) when the customers use waste heat generated by the fuel cell. These systems have an initial electrical efficiency greater than 40% and an average lifetime electrical efficiency of 38%. Utilization of all of the waste heat generated by the system allows the customer to achieve 90% overall utilization.

## Phosphoric Acid Fuel Cells

PAFCs are the first fuel cells to be commercially available. The major manufacturers of these fuel cells are UTC Power, Toshiba Corporation, HydroGen Corporation, Fuji Electric Corporation and Mitsubishi Electric Corporation. UTC Power introduced for sale a 200 kW PAFC system in 1991, and over 260 units were delivered to various customers worldwide. The design operational lifetime for these units was 40,000 h and most of the fielded units have met or exceeded this requirement. A number of these units are still operational today with fleet leader at Mohegun Sun in Uncasville, Connecticut, USA, accumulating more than 76,000 h [48]. Fuji's phosphoric acid fuel cell power plants, launched in 1998 have also demonstrated 40,000 h of life in field and some units after overhaul have exceeded 77,000 h of operational lifetime [1].

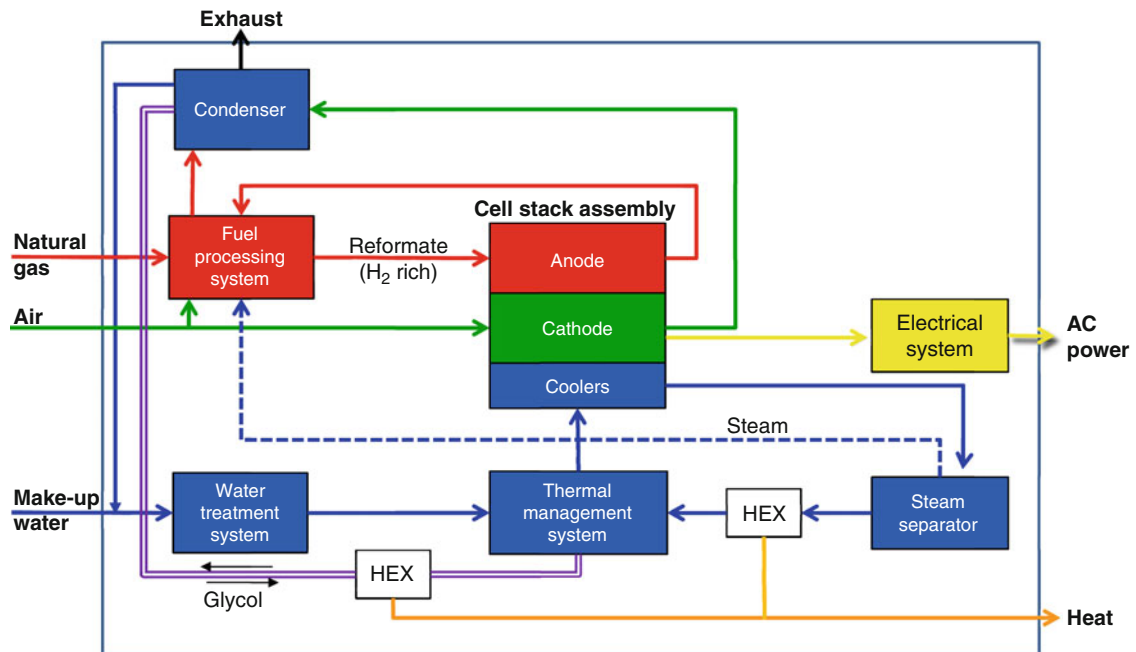
Phosphoric acid fuel cells usually operate on natural gas but they can operate on other fuels such as H<sub>2</sub> exhaust from chemical plants and anaerobic digester gas (ADG) from waste treatment plants. UTC Power's

phosphoric acid fuel cell system is usually designed to operate in water balance, i.e., it does not consume water from the site nor produce excess water at the site. These power plants operate at ambient pressures and have the capability to transition between grid connect and grid independent model. This capability allows the customer to draw electricity when required from grid and export excess electricity generated by the fuel cell power plant when the customer loads are lower than the power generated by fuel cell. Basic description of an ambient pressure natural gas operating UTC Power's PAFC system is shown in Fig. 1.

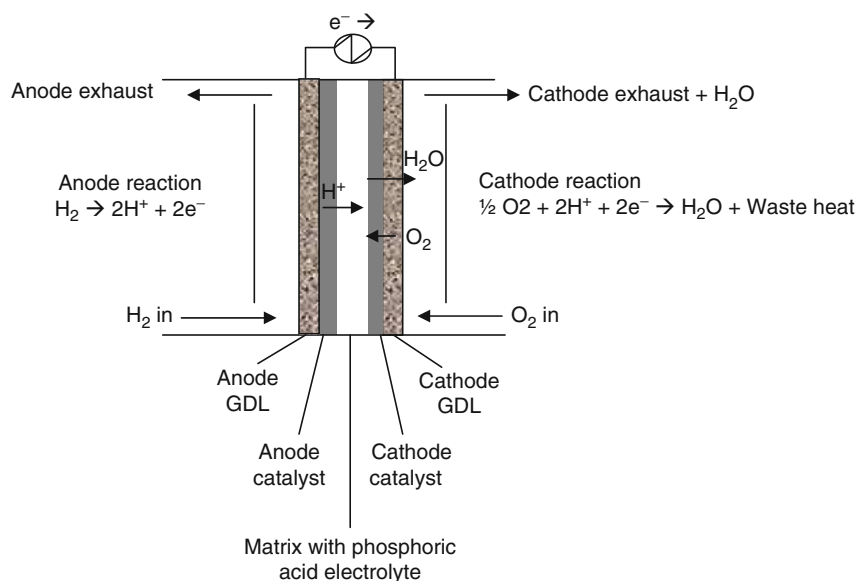
Natural gas is fed to the fuel processor where  $\text{CH}_4$  in natural gas is converted to  $\text{H}_2$ -rich fuel using steam reforming.  $\text{H}_2$ -rich fuel from reformer is fed to the cell stack where it is electrochemically combined with  $\text{O}_2$  in air to generate DC power and thermal energy. DC power is sent to power conditioning system where it is converted to AC power. Makeup water is used only during the start-up of power plant to fill the thermal management system or in situations where the power plant is operated outside its design specifications for extended period of time.

The cell stack assembly is the heart of the power plant which produces both electrical and thermal energy by electrochemically combining  $\text{H}_2$  in fuel and  $\text{O}_2$  in air. The basic description of a phosphoric acid fuel cell is shown in Fig. 2.

PAFCs operate at temperatures between  $150^\circ\text{C}$  and  $225^\circ\text{C}$ .  $\text{H}_2$  from fuel is split into protons and electrons at the anode electrode. The protons travel through the electrolyte ( $\text{H}_3\text{PO}_4$ ) and reach the cathode catalyst layer where they combine with  $\text{O}_2$  in air producing DC power, water, and waste heat. The electrolyte is usually held in a refractory nonconducting matrix like silicon carbide. Water generated in the fuel cell is removed by cathode exhaust. Cathode exhaust is sent to the condenser where generated water is condensed. This condensed water is sent back to the water treatment system where it is purified to minimize the conductivity and sent to the thermal management system. Coolant exhaust from cell stack is sent to steam separator where steam required for fuel processor is separated and the remaining hot coolant water is sent through multiple heat exchangers to supply thermal energy to customers.



**Phosphoric Acid Fuel Cells for Stationary Applications. Figure 1**  
Description of an ambient pressure operating PAFC system

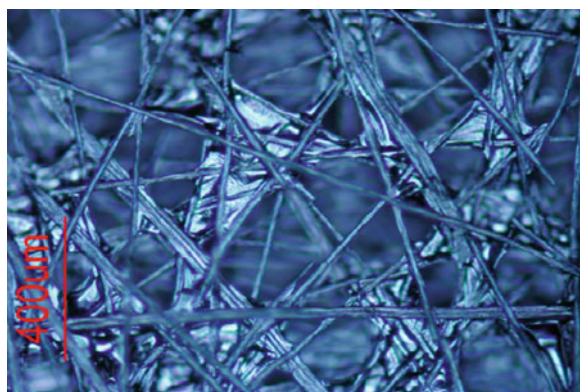


**Phosphoric Acid Fuel Cells for Stationary Applications. Figure 2**  
Phosphoric acid fuel cell

### Cell Stack Assembly Components

The main components of a PAFC are bipolar plates, gas diffusion layers, catalyst layers, and matrix layer [3]. Typical cell designs include a sandwich of these layers as arranged in Fig. 2 between coolers. Multiple cells per cooler designs are generally employed to improve power density and this unit is called a sub stack. A cell stack assembly consists of multiple sub stacks held between two pressure plates under compression to minimize reactant leakage and contact resistance losses.

*Gas diffusion layers (GDLs):* As the name suggests, GDL allows for gas to diffuse from the bulk flow in channels to the catalyst layer; provides mechanical support to the catalyst layer; allows for water management, i.e., removal of product water to the bulk flow in channels and heat transfer from the catalyst layer to the coolers. GDLs are made by turning polyacrylonitrile (PAN)-based carbon fiber into carbon paper. Carbon paper is then impregnated with a phenolic resin by a prepreg process, hot laminated to cure the resin in place and to achieve the desired thickness. It is then heat treated at  $\sim 1,000^{\circ}\text{C}$  to turn the phenolic resin into carbon and then graphitized at  $\sim 2,500^{\circ}\text{C}$  in an inert atmosphere [4, 5]. Graphitization helps impart better thermal and electrical conductivity to the GDL.



**Phosphoric Acid Fuel Cells for Stationary Applications. Figure 3**  
Graphitized carbon fibers in PAFC GDL

A finished substrate is approximately 70% porous with a mean pore size of 25–30  $\mu\text{m}$ . These finished substrates as shown in Fig. 3 are usually hydrophobic and hence they are impregnated with a wettability coating (e.g., Vulcan or Black Pearl Carbon) to facilitate water and acid management.

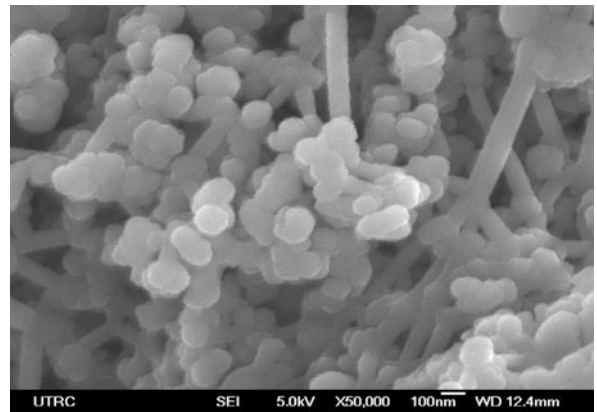
*Electrolyte matrix:* Matrix in PAFC holds  $\text{H}_3\text{PO}_4$  electrolyte and hence facilitates the movement of protons from anode to cathode. The matrix layer should be

wettable, sufficiently porous to hold  $\text{H}_3\text{PO}_4$  electrolyte, should have minimal reactivity with  $\text{H}_3\text{PO}_4$  during operating temperatures, have sufficient electrical isolation to prevent shorting in cells and finally have sufficient bubble pressure ( $\sim 35$  kPa) to minimize reactant crossover. Materials such as wettable PTFE, papers formed using organic polymers and silicon carbide have been evaluated as matrix layers by various fuel cell manufacturers. Wettable PTFE layers failed because they lost their wettable properties over the course of operation leading to expulsion of acid from the matrix layer and hence cell failure due to reactant crossover. Organic polymers such as polyetherketones (PEK) and polybenzimidazoles (PBI) have acceptable beginning-of-life properties but seem to lose desirable properties over the course of time. Silicon carbide matrix layer has been used in UTC Power's PAFC and it has retained its material properties over long life operation. This matrix layer is formed by spraying  $5\ \mu\text{m}$  particle size SiC with 5% PTFE as binder onto the catalyst layer and then heating the coated electrode to  $\sim 300^\circ\text{C}$  to evaporate the solvent and form the layer. These layers are typically  $25\text{--}50\ \mu\text{m}$  to minimize ionic losses, 50% porous, have a bubble pressure of  $\sim 70$  kPa and an effective ionic resistivity of  $6\text{--}7\ \Omega\text{cm}$  [6, 7]. They have worked pretty well in UTC Power's PureCell<sup>®</sup> Model 200 for greater than 70,000 h.

*Cathode catalyst layer.* Cathode catalyst layers in PAFCs are made using a mixture of PTFE and catalyst-coated graphitized carbon. These layers are around  $100\text{--}125\ \mu\text{m}$  thick and are approximately 70% porous. Catalyst is usually Pt or an alloy of Pt. Graphitized carbon is used on cathode to ensure minimal carbon corrosion during long life operation. Oxidizing environment on cathode can cause significant corrosion issues during steady state operation and starting/stopping of the power plant. Thus, to minimize carbon corrosion for long life (10–20 years), it is essential to use corrosion-resistant graphitic carbon [8, 9]. PTFE is used in the mixture to produce an optimum balance between hydrophilic and hydrophobic pores in the catalyst layer. Hydrophilic pores take up acid and allow for proton transport while hydrophobic pores allow for reactant air or fuel to reach catalyst sites. It is very critical to achieve this balance to have an optimum electrolyte fill in the catalyst layer while providing enough path ways for gas to reach the catalyst.

If the catalyst layers are underfilled or have low electrolyte take-up, it results in the cell having high IR losses; if the catalyst layers are overfilled or have high electrolyte take-up, it results in low cell performance due to flooding or mass transport losses. Hence it is very critical to achieve optimum fill level and remain at that fill level during the course of operation. UTC's catalyst layers have demonstrated the ability to retain this pore structure over the course of field operation. UTC's cathode catalyst layer is manufactured in batch process [10–13]. Graphitized carbon coated with Pt alloy is mixed with PTFE particles to form floc using various wet mixing techniques. This floc is then dewatered and dried to form floc pellets. Floc pellets are then finely ground and deposited onto the GDL using a cloud tower. These electrodes then go through a sintering oven where the PTFE flows over the graphitized carbon to form hydrophilic and hydrophobic pores. SEM image of a typical PAFC electrode where catalyst-coated carbon particles are further coated with PTFE is shown in Fig. 4.

*Anode catalyst layer.* Anode catalyst layers are manufactured using the same process described above but these catalyst layers use carbon instead of graphitized carbon. Since anode electrode is mostly in  $\text{H}_2$  environment, there is very little to no corrosion of the anode electrode and hence the use of carbon instead of graphitized carbon [9]. Regular carbon has a higher surface area than graphitized carbon thus allowing for better Pt loading characteristics and thus improved



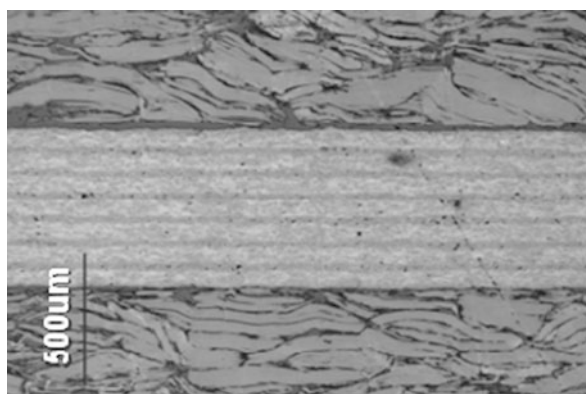
Phosphoric Acid Fuel Cells for Stationary Applications.

Figure 4

PAFC catalyst particles coated with PTFE

H<sub>2</sub> oxidation reaction kinetics. Anode catalyst layers are made slightly more hydrophilic than cathode catalyst layers since H<sub>2</sub> kinetics are fast and thus mass transport losses set in at much higher current densities than operating current densities.

**Bipolar plates:** The main function of bipolar plates is to transport reactants to the catalyst layers, prevent reactant crossover from one cell to another, prevent acid migration from one cell to another, provide good electrical properties to minimize IR loss through the plate and provide good thermal properties to transport heat generated in the cell to the coolers. UTC Power's bipolar plates are usually made of mixture of flaky graphite and FEP. FEP is limited to less than 20% to minimize the impact of IR loss and thermal conductivity loss in the plate. This mixture of FEP and graphite is placed in mold and compacted at high pressures (500–1,000 psi) to make molded preforms. The bipolar plates in PAFCs should have very low porosity to reduce electrolyte take-up. Reducing the electrolyte take-up prevents the formation of a continuous electrolyte pathway through the thickness of the plate thereby mitigating electrolyte pumping as discussed in electrolyte management section. In order to achieve this low porosity, two preforms are laminated on either side of a highly nonporous separator plate to form the integral separator plate as shown in Fig. 5. After lamination of molded preforms to the separator plate, channels are either machined or molded in the plate. Separator plate is manufactured by impregnating



Phosphoric Acid Fuel Cells for Stationary Applications.

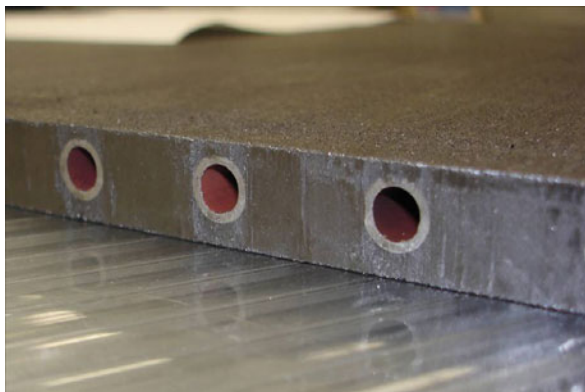
Figure 5

UTC's phosphoric acid fuel cell bipolar plate

carbon paper with a phenolic resin. After resin impregnation, multiple layers of paper are laminated and carbonized in a slow carbonization process at 1,000°C. The carbonization process has to be slow to prevent the formation of pores in the structure. After carbonization, the plate is graphitized at ~2,500°C to improve corrosion resistance, thermal and electrical properties. UTC Power's bipolar plates have <5% porosity with very high tortuosity and have performed robustly in the field in mitigating acid transfer through plane in the plate. Channels in these plates are coated with wettable carbon to allow for acid management during fuel cell operation [14]. The perimeter of the plate also needs to be hydrophobic to prevent electrolyte being pumped on the edge of the plate. This hydrophobic break is achieved by placing PTFE flaps around the perimeter [15–22].

**Edge seals:** UTC Power's cell stack design uses external manifolds to transport reactants in and out of the cell stack. As a result, the edge of the anode GDL in air manifold and similarly cathode GDL in fuel manifold need to be sealed against reactant mixing in these locations. The obvious method of sealing is to fill the perimeter of the GDL with fuel cell-compatible filler material to minimize porosity. UTC Power's design fills the perimeter with a wettable carbon. Some of the previous designs also accomplished edge sealing using silicon carbide. The particle size of this filler material is usually around 5 µm and thus when impregnated into the GDL it forms pores that are much smaller than the pores in the GDL. Edge seals are made by forming a viscous ink of these particles and using a screen printing process to apply the ink to the GDL. Subsequent drying of the electrode removes the volatiles used in the ink leaving carbon in the GDL. When these GDLs are filled with acid during the assembly process, the edge seals take up acid due to capillary action and thus form wet seals which prevent gas from leaking into manifolds [23–26].

**Coolers:** Coolers in UTC Power's PAFCs are made of stainless steel tubes embedded in molded material (mixture of FEP and Graphite) as shown in Fig. 6. These coolers use the same manufacturing process as that of the bipolar plates. A serpentine cooler tube is placed between two molded preforms and laminated together at high pressures and temperatures. The ends of the tubes coming out of the molded material are



### Phosphoric Acid Fuel Cells for Stationary Applications.

#### Figure 6

UTC Power's PAFC cooler

wrapped in PTFE film or coating to prevent attack of stainless steel tubes by phosphoric acid. In UTC Power's PAFC system, water is used as coolant. This coolant enters the cell stack as liquid water and exists as a mixture of steam and liquid water between 150°C and 180°C. As a result, the coolant temperature rises as the water is heating up in single phase and once it reaches the saturated pressure, it stays at the same temperature but starts generating steam inside the coolers. The steam from coolant exit is used in the reforming reaction and the remaining hot water is used to supply thermal energy to customers. Due to the two-phase cooling, it is very essential to ensure that the liquid pressure drop is sufficiently greater than the two-phase pressure drop. If not, the variability in performance between cells can cause the system to run into thermal imbalance and as a result some cells would be operating much hotter than other cells, decreasing their life substantially. Increase in liquid pressure drop is obtained by inserting orifices or some other kind of flow restrictors at the coolant inlet to the cell stack. Orifice design has to ensure that there is no potential for clogging of coolant inlet to the cell stack. Thus, in order to mitigate issues associated with clogging, increase in pressure drop is accomplished by using a long tube coiled in front of each cooler [27–30].

*Non-repeat components:* The components described in the previous section are usually referred to as repeat components since every cell has those components with multiple cells present in a cell stack. Non-repeat components are those that are used only once in a cell

stack. The main non-repeat components in a cell stack assembly are pressure plates, coolant inlet and outlet manifolds, reactant inlet and outlet manifolds, and manifold seals. Multiple cells placed between coolers are stacked between two stainless steel pressure plates and loaded axially to around 60 psi with the help of tie rods that run the entire length of the cell stack. Reactant manifolds are then assembled onto the cell stack with manifold seals placed between the cell stack and the manifold. These reactant manifolds are made of stainless steel and coated with PTFE coating to prevent phosphoric acid attack of the manifolds. It is very essential to ensure that there are no pin holes in the PTFE coating for the same reason mentioned above. In phosphoric acid fuel cells, manifold seals made with high fluorine content fluoroelastomers. Fluoroelastomers are very resistant to hot phosphoric acid environment and hence they have lifetimes greater than 10 years. It is very essential that these seals have very low porosity. Seals with high porosity take up acid and due to the potential difference between the top and bottom of the cell stack, acid pumps to the top of the stack thus causing the bottom of the stack to fail due to loss of acid and top of the stack to fail due to flooding by acid. In addition, these seals need to conform to the variations in the layout of cells along the height of the stack. UTC Power uses a seal mechanism where a cured fluoroelastomer is placed adjacent to the manifold and uncured fluoroelastomer placed adjacent to the cell stack. During the heat-up of the cell stack to operating temperature, the uncured manifold seal melts, flows, and cures in place to seal the skyline that is formed due to the variations in the layout of cells.

### Cell Stack Assembly Life

Phosphoric acid fuel cells life is primarily a function of catalyst decay and acid management.

*Catalyst decay:* Catalyst decay occurs due to steady state operation and start-stops. Steady state decay occurs due to agglomeration of Pt particles in the catalyst layer. Pt deposited onto the carbon support in PAFC catalyst has a particle size of approximately 4–5 nm. Due to the high temperature of operation and high operating lifetimes (~80,000 – 100,000 h), these Pt particles tend to agglomerate and loose surface area. Smaller the particle size, higher the surface area and

more sites for  $O_2$  reduction or  $H_2$  oxidation, but these particles are much more unstable and tend to agglomerate quickly. Even though these Pt catalysts tend to have high performance at beginning of life, their performance drops quickly due to Pt particle agglomeration. UTC Power's PAFCs use a Pt alloy, Pt–Cr–Co on cathode and Pt on anode. The use of alloy on cathode allows for improving the surface area while reducing the agglomeration of catalyst. On anode pure Pt is used with approximately one third loading of that of the cathode. Low loadings are used on anode since  $H_2$  kinetics are fast. These Pt particles are approximately 2–3 nm in size and tend to agglomerate faster than the cathode. As a result, anode catalyst electrochemical area decreases much faster than cathode.

UTC Power has deployed more than 260 Purecell<sup>®</sup> model 200 power plants and their performance decay was as expected in field. This model power plant operated at 200 kW and Fig. 7 shows the operational experience of this model.

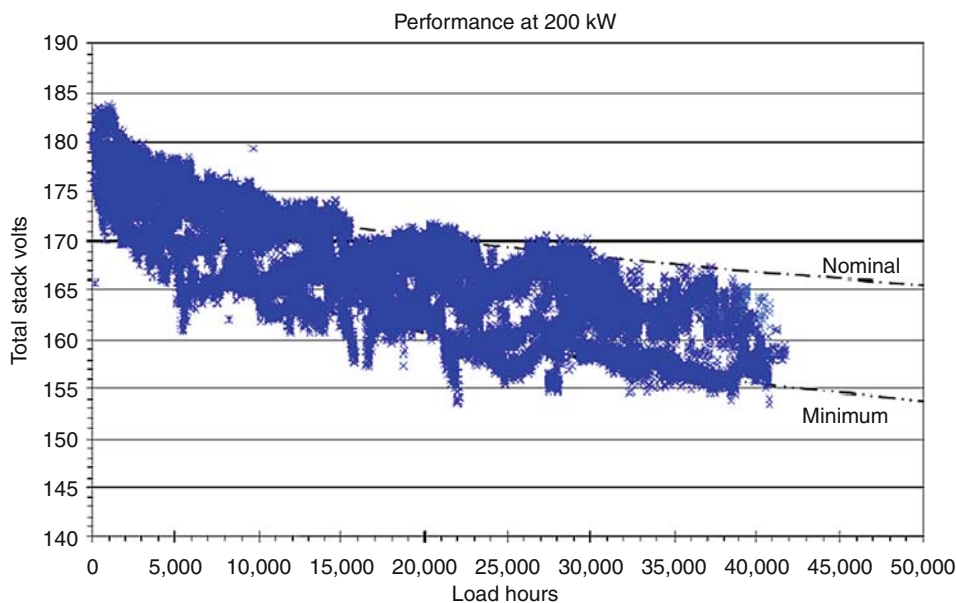
Teardown analysis has been performed on some of these units that have returned from field after their design operational lifetime of 40,000 h. The catalyst in these power plants has degraded as mentioned above. The average ECA of a new catalyst sample is  $50 \text{ m}^2/\text{g}$

while catalyst that has aged in field for 43,000 h has an average ECA of  $6.5 \text{ m}^2/\text{g}$ . Field operation increased the particle size from 4–5 nm to 20–25 nm. Figure 8 shows how catalyst ages with operational time.

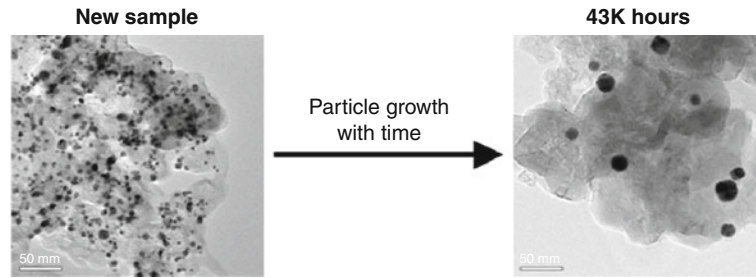
Based on Model 200 experience, UTC Power has developed model 400 which has a design life of 85,000 h or 10 years. To meet 10 year performance requirements, cell operating conditions have been changed in addition to modification of catalyst properties during manufacture.

In addition to improving the catalyst, system strategies such as using a mixture of  $H_2/N_2$  as purge gas during shutdown and use of voltage clipping are being deployed to mitigate start/stop losses in model 400. Figure 9 shows performance data of model 400 catalyst vs. design requirements.

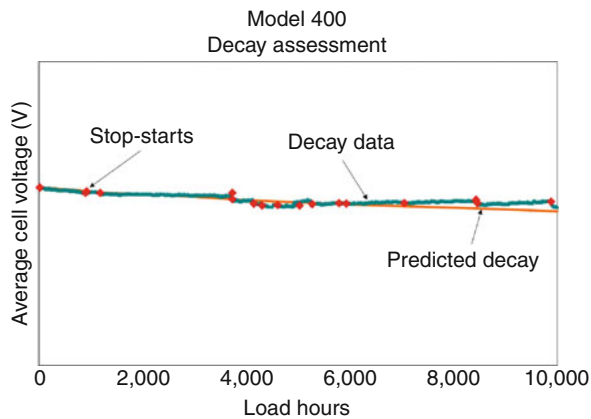
*Electrolyte management:* Phosphoric acid has a finite vapor pressure and a low contact angle with all of the components used in fuel cells except for catalyst layers. As a result, managing evaporation and migration of acid is very critical to maintaining the life of the fuel cell. Evaporation of acid from a cell is controlled by the reactant exhaust temperature. If the reactant exhaust temperature is high, the acid lost in vapor phase increases. Acid loss due to temperature is



Phosphoric Acid Fuel Cells for Stationary Applications. Figure 7  
Purecell<sup>®</sup> model 200 operational performance



Phosphoric Acid Fuel Cells for Stationary Applications. Figure 8  
PAFC catalyst on carbon



Phosphoric Acid Fuel Cells for Stationary Applications.  
Figure 9  
PureCell<sup>®</sup> model 400 performance verification

exponential in the region of operation of phosphoric acid fuel cells. In addition, reactant flow has a linear relationship to acid loss. Hence, if the reactant utilization in a cell is low, the loss of acid is high due to higher exit flow. PAFC cell designs need to balance the loss of acid due to exit temperature and flow. UTC's cell designs operate such that the cell is hot enough to ensure maximum performance with high reactant utilization while the reactant exhaust is cold enough to ensure long life operation [31–33]. This is accomplished by placing coolant inlet to the cell near the reactant exhaust. The cell adjacent to the cooler is the coldest cell and the cell equidistant to coolers on either side is the hottest. Cold cell acid loss is less than hot cell acid loss, and thus hot cell is the life limiter due to evaporation. Water management is not an issue in phosphoric acid fuel cells. Design features such as

incorporating a nonactive condensation zone near the exit location of the reactant stream, having an electrolyte reservoir in the cell, etc., are also used in addition to managing the thermal profile of the cell. UTC Power's phosphoric acid fuel cells operate at ambient pressures and between temperatures of 150°C and 225°C. As a result, water generated in the fuel cell is in vapor phase and leaves with the reactant exhaust. A PAFC cell is generally in water balance. These cells operate between 95% and 105% acid concentration depending on cell operating conditions. If the reactant flow is high, the acid in the cell is more concentrated and more water vapor leaves the cell and vice versa if the flow is low [3].

The movement of acid within the cell is managed using capillary action between various layers. The GDLs have the biggest pore size followed by matrix layer and finally the catalyst layers. During operation, acid is lost due to evaporation from the GDLs as they have the lowest capillary pressure of all the components. Once GDLs reach ~2–3% fill level, the pores in the matrix layer start emptying. Void spots in the matrix layer cause gas crossover between anode and cathode resulting in cell failure.

Acid movement within a cell is a complex phenomenon. During the build of a phosphoric acid fuel cell, acid is deposited onto the porous components [34, 35]. When load is applied to the cell, protons start moving from anode to cathode. Since phosphoric acid is a weak acid, it dissociates and is in the form of  $H^+$  and  $H_2PO_4^-$  in the cell. Thus, in order to maintain charge balance, these phosphate ions start moving from cathode to anode. This flow of acid from cathode to anode is balanced by the liquid pressure difference between anode and cathode. Once acid starts accumulating in



the anode, liquid pressure in the anode GDL increases and starts balancing the flow of acid due to charge imbalance. Thus, at steady state, approximately 75% of the acid deposited into the cathode GDL during build moves into the anode GDL. Evaporation over the course of life results in depleting anode and cathode GDL fill levels, and at end of life, matrix layer starts emptying of acid resulting in reactant crossover.

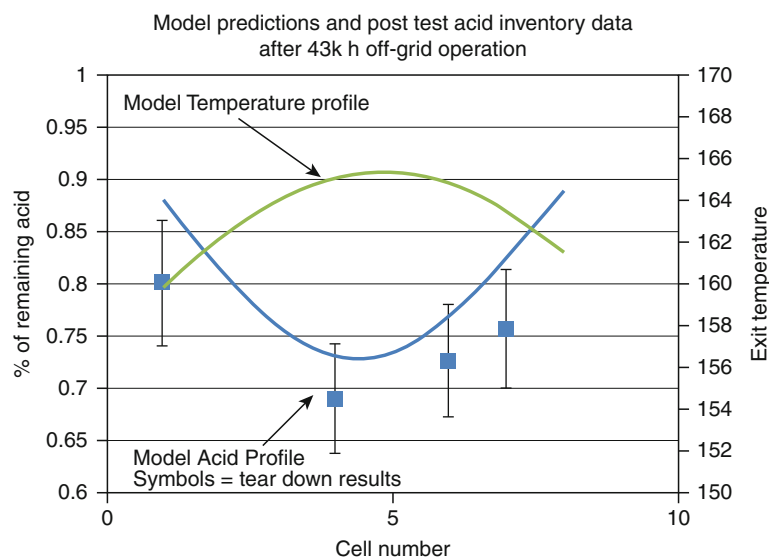
Acid can also move between cells if appropriate hydrophobic breaks are not inserted between cells [16, 17]. Bipolar plates have very low porosity but they are easily wetted by acid. As a result, the edge of the bipolar plate gets wetted by acid easily and forms a liquid connection between two adjacent cells (referred to as cell 1 and cell 2) across the bipolar plate. In a typical cell one side of the bipolar plate is in reducing environment ( $H_2$  flow channels) and the other side of the plate is in an oxidizing environment (air flow channels). In a simple explanation, electrolyte in cathode GDL of cell 1 is near the cathode potential of the cell 1. The electrolyte potential across this bipolar plate between cell 1 and cell 2 is close to the hydrogen reference potential. As a result, the potential difference across the bipolar plate is equal to that of one cell's voltage. This potential difference can drive a very small shunt current through the acid film that has formed on the edge of the plate thus moving protons from cell 2 to

cell 1. Due to movement of protons from cell 2 to cell 1, phosphate ions start moving from cell 1 to cell 2. Over the course of short time intervals, this shunt current can drive enough acid such that cell 1 fails due to dry out and cell 2 fails due to flooding by acid. UTC's cell design includes a hydrophobic break or PTFE flap between cells to mitigate the formation of a continuous film between cells thus mitigating movement of acid due to shunt currents [16, 17].

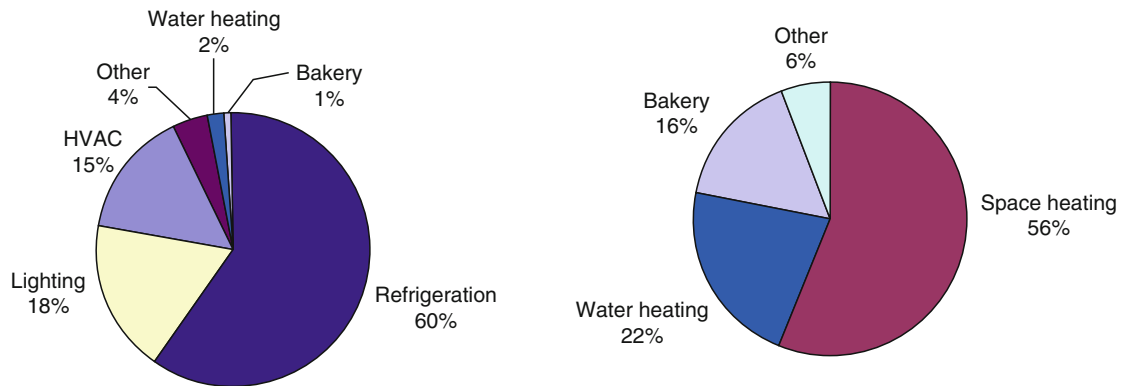
Teardown of UTC's Purecell<sup>®</sup> Model 200 units from field and measuring the acid content in these cells indicates that the PTFE flap between cells has worked very effectively in mitigating shunt migration of acid between cells. In addition, the loss of acid due to evaporation is very close to model predictions [36] as shown in Fig. 10.

### PAFC Applications

PAFC fuel cells are a natural fit for combined heat and power applications. The use of CHP fuel cell systems in commercial buildings such as supermarkets, office towers, schools, data centers, industrial buildings, etc., improves overall efficiency by displacing low efficiency electricity provided by grid while providing enough thermal energy to displace fuel required for space heating and/or domestic hot water.



**Phosphoric Acid Fuel Cells for Stationary Applications. Figure 10**  
Acid loss as a function of cell temperature



**Phosphoric Acid Fuel Cells for Stationary Applications. Figure 11**  
Supermarket energy consumption [37, 38]

UTC's PAFC system has been used in a wide variety of applications and given below is an example of how this application can be used in supermarkets. A typical supermarket's energy needs are met 80% by electricity and remaining by natural gas. For a typical supermarket it has been estimated that a 10% reduction in energy costs is equivalent to increasing net profit margins by 16%. In other words, \$1 in energy savings is equivalent to increasing sales by \$59 [37, 38]. Hence energy-efficient methods of operating stores are a top priority for supermarkets. Supermarket's energy usage is primarily a function of the square footage of the store and its operating hours. Electricity in a supermarket is used mostly for refrigeration while natural gas is used mostly for space heating as shown in Fig. 11.

Supermarkets need reliable and low-cost energy to maintain freshness of produce and to improve their margins. Using a fuel cell application with CHP capabilities can improve their efficiencies significantly thereby reducing energy costs. UTC Power's PAFC system is being currently evaluated by various supermarkets to achieve high efficiency and hence reduced energy costs. Shown in Fig. 12 is a Model 400 application which supplies the store with 324 kW of electricity, 63 kW of high-grade heat which is used for space heating and refrigeration using absorption chilling and 92 kW of low-grade heat for domestic hot water. Absorption chilling is driven by heat energy rather than mechanical energy [38, 39]. The power plant uses 750 kW (LHV) of natural gas to generate above the energy required by the store thus achieving overall efficiency of

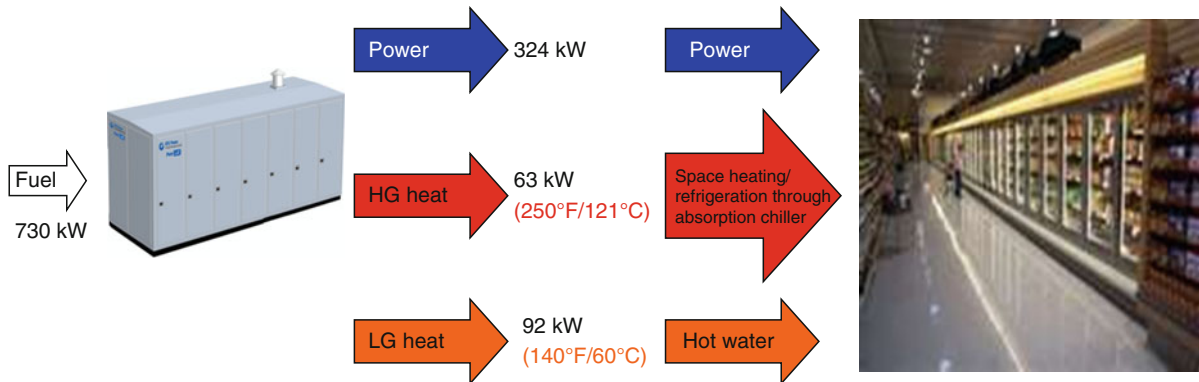
65%. As energy costs continue to increase, CHP applications provide the value proposition required by customers to improve their margins.

Another example of UTC's Purecell® model 200 fuel cell system being used for commercial application is the installation at Mohegun Sun Casino in Uncasville, Connecticut, USA. This facility uses both high-grade and low-grade heat from the fuel cell along with electric power (Fig. 13). Customer needs heating all year long and effective integration allowed for achievement of ~85% efficiency with this unit [38, 39].

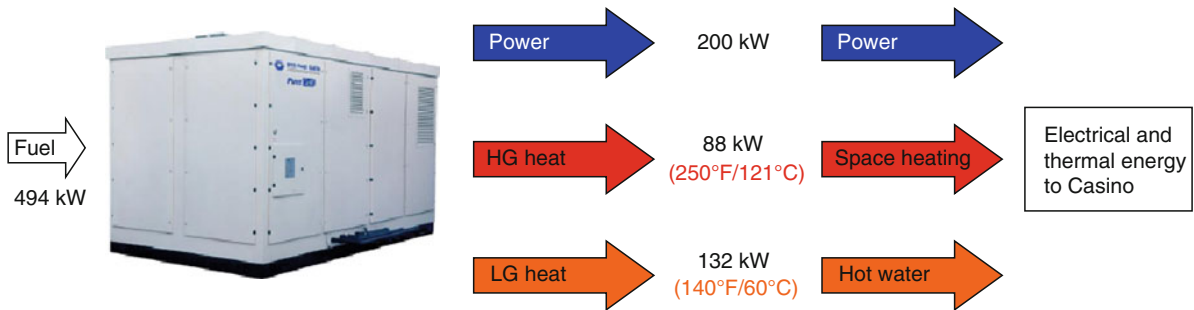
### Future Directions

Cost must be reduced aggressively to enable commercialization of combined heat and power (CHP) fuel cells. A significant portion of the fuel cell system cost is the cost of cell stack materials and manufacturing processes including labor. Cost of fuel cell components is roughly split half and half between materials and manufacturing costs. Significant R&D investments to enable continuous efficiency improvements along with operational excellence are essential to drive down manufacturing costs. Development of high-volume manufacturing techniques along with high-speed quality control process will enable reduction of a significant portion of this cost [40].

Unique processes used in the manufacture of PAFC components result in increased part cost. Current processes for manufacturing GDLs, electrodes, bipolar plates, etc., employ batch processes and require



**Phosphoric Acid Fuel Cells for Stationary Applications. Figure 12**  
Use of a PAFC system for supermarket application



$$\text{Maximum efficiency} = \left( \frac{200+88+132}{494} \right) = 85\%$$

**Phosphoric Acid Fuel Cells for Stationary Applications. Figure 13**  
Use of a PAFC system for commercial building application

continuous manufacturing processes to reduce cost and also to meet projected high volumes.

GDL manufacturing for high-temperature CHP fuel cells currently involves a lot of batch processes such as turning the fiber into resin-impregnated felt, carbonization, graphitization (for tolerance to high-temperature operation), etc., before electrodes are deposited onto the part. Continuous manufacturing processes for making GDLs can be enabled by use of double belt press, but this requires process development. Double belt press process integrates all of the various steps involved in making GDLs wherein fiber

can be fed from one end of the press with finished substrate emerging from the other end of the press. Similarly, bipolar plates used in PAFCs are currently manufactured using unconfined compression molding followed by machining, wet treatment of bipolar plate and incorporation of hydrophobic edges to prevent acid migration from cell to cell. Continuous manufacturing processes such as net-shaped molding can enable integrating these operations into one operation thus enabling high-volume manufacturing and significantly reducing the cost of component. Finally, fuel cell industry needs to move toward manufacturing

that can help bring the processes to the part rather than parts moving through the process to reduce factory footprint costs, handling costs, and energy costs, and thus reduce cost of manufacturing. Fuel cell industry currently adopts processes that have been developed in other industries. For example, Gravure coating which is used to coat thin layers at high speeds is being used to coat thick layers of electrolyte matrix at low speeds. The fuel cell industry needs the development of new manufacturing processes for fuel cells that are tailored toward optimum manufacture of a part keeping its fundamental requirements in consideration.

Since every component used in a cell (GDL, catalyst layer, bipolar plate, matrix layer, etc.) is critical for the operation of the cell stack, these components have to be within a very tight specification and meet defined key product characteristics. This requires the implementation of a robust quality control system. For example, substrates in PAFCs have to be checked for porosity, IR, thermal conductivity, thickness, density, compressive strength, flex strength, etc. All of these quality checks impart cost to the substrate. Ability to integrate these measurements into one tool will enable help high-volume manufacturing. Similarly high-speed measurement techniques (e.g., ultrasonic crack detection methods) that can detect defects/cracks in electrode layers and bipolar plates in a few seconds are required for high-volume manufacturing of these components. In addition, improving manufacturing processes to achieve high level of robustness will reduce the number of cell stack rebuilds and/or infant mortalities thus reducing cost. Finally, all of the proposed high-volume manufacturing techniques need to yield product at a very minimal scrap rate to reduce part cost.

Another approach for reducing cost is to improve the efficiency of the power plant. Since cell stack has the lowest efficiency of all the components, improving its efficiency provides maximum benefit but at the same time cell efficiency improvements are not easy to achieve. Improved efficiency can be capitalized by removing cells out of the cell stack to generate the same power at baseline efficiency or it can be used to provide cheaper power to the customer. Phosphate ions poison cathode catalyst, i.e., they occupy valuable sites on Pt catalyst thus reducing the number of sites

available for oxygen reduction reaction. This leads to lower cell performance (typically  $< 0.25 \text{ W/cm}^2$  of electrode) and hence more number of cells making the stack cost challenging. The only way to eliminate phosphate poisoning is to use a new electrolyte instead of  $\text{H}_3\text{PO}_4$ . New liquid electrolytes which do not have phosphate poisoning effect can lead to significantly improved electrical efficiency. Truls Norby's paper on "Solid state protonic conductors: Principles, properties, progress and prospects" [41] discusses properties of potential materials that can be used as electrolytes in the temperature ranges discussed. Another way of improving efficiency could be by improving the  $\text{O}_2$  solubility/diffusion in  $\text{H}_3\text{PO}_4$ . Literature shows that additives to  $\text{H}_3\text{PO}_4$  can improve  $\text{O}_2$  solubility in  $\text{H}_3\text{PO}_4$  and hence improved efficiencies. Additives such as silicon oils, C4 and C6 compounds, Pyrroles, protonated polyamine, etc., evaluated by various researchers have shown performance improvements but further research is needed to determine if these performance improvements are sustainable [42].

Conditioning and qualification of cell stack after it is assembled and final qualification of power plant after integrating cell stacks with balance of plant (BOP) takes significant time increasing product cost. Conditioning of cell stacks involves processes that cure seals and set them in place and enable electrolyte movement into the catalyst layers making them functional. After conditioning, acceptance testing is performed on cell stacks where diagnostics are done to ensure that flow to the cells is appropriate, seals are set in place, catalyst layers are functional, and that the product meets design performance requirements. Ability to assemble hundreds of cells in a power plant in the same manner every time will enable the industry to go to a limited sampling technique rather than testing every power plant thus reducing costs. Developing methods to perform conditioning and acceptance testing on modules of cells rather than on fully assembled cell stack will enable removal of problem parts early in the build and prevent costly teardowns. Further understanding of the electrode manufacturing process at a fundamental level could enable integrating conditioning procedures into manufacturing process thus eliminating post component manufacture conditioning and acceptance testing. R&D aspects that can reduce or eliminate conditioning

and acceptance testing of CHP fuel cells should be developed.

Balance of plant (BOP) in fuel cell power plants also adds significant cost to the power plant. The ability to deliver contaminant-free fuel to the cell stacks, ability to provide reactants very quickly as the load changes, etc., add significant cost to the power plant. The fuel processing system (FPS) is one of the most expensive subsystems in the fuel cell power plant BOP, at 20% of total power plant cost. The catalytic steam reformer (CSR) is the single most expensive component (18%) within the FPS. The CSR is effectively a catalyst-augmented heat exchanger. As such its cost can be significantly reduced by manufacturing designs and techniques to improve heat transfer such as direct application of catalyst to reformer walls (catalyzed walls improve heat transfer by eliminate film losses, thus enabling reduction heat transfer area), improve catalysts to reduce poisons to the stack (such as and sulfur and ammonia which accelerate stack decay and increase stack costs), improve high-volume manufacturing techniques such as spin casting to lower tube costs, and advanced casting techniques to form the reformer burner which is presently made by welding many smaller sheet metal pieces. Automated welding processes are also critical to obtaining repeatable high-quality welds at high volume and low cost. A significant but overlooked BOP cost is piping. Cost reductions can be realized by replacing complex pipe runs requiring threaded or welded assembly with pipe or tube bending. Improved pipe bending capability can reduce manufacturing and assembly time and expense, especially at high volumes.

Heat exchangers are essential in fuel cell power plants to control temperatures for the major components (stack, reformer, shift converter), provide valuable cogeneration energy to customers, and reject waste heat. Brazed plate heat exchangers offer the most cost-effective heat transfer in many applications, but are often limited by poor thermal cycle durability which effects transient capability, especially in CHP power plants where load following and variable customer heat usage cause numerous thermal cycles. Manufacturing development to improve the durability of the brazed plate-type heat exchangers would significantly reduce overall BOP costs by enabling their use

in more applications within the CHP fuel cell power plants.

Another critical area of the BOP components is the power conversion system as this system is the customer interface to the fuel cell and contains much of the key product performance characteristics necessary to achieve the customers' needs. Significant cost reduction could be achieved by developing a modular high-frequency power conversion system. A modular approach would enable use of the same design across the entire product line offered to various customers. Using this same design would allow for improvements in manufacturing volumes as the manufacturing of all current power conversion designs could be consolidated instead of the current mix of various designs in low volume. This would also drive down the cost of the power conversion system by increasing the volume of similar components that would be sourced for the power conversion of the various products offered. Having a modular design would also reduce test time in manufacturing as this would enable improved testing techniques in the supply base (i.e., automated testing) and reduce the mix of power conversion systems encountered at the final assembly.

Development of markets at lower volumes of 50–100 units/year needs to be facilitated so that design-driven cost reduction can be learned out. Demonstration programs where state or federal commercial buildings are converted to using power from fuel cell power plants should be encouraged. Finally, cost reduction as a key aspect of demonstration programs should be emphasized.

### Acknowledgments

The authors would like to acknowledge Tom Jarvi for his valuable input into framing the outline for the entry.

### Bibliography

1. [www.fujielectric.com](http://www.fujielectric.com)
2. [http://www1.eere.energy.gov/hydrogenandfuelcells/fuelcells/fc\\_types.html](http://www1.eere.energy.gov/hydrogenandfuelcells/fuelcells/fc_types.html)
3. Breault RD (2003) Stack materials and stack design. In: Vielstich W, Lamm A, Gasteiger HA (eds) Handbook of fuel cells: fundamentals, technology and applications, vol 4, Part 4. Wiley, Chichester, pp 797–810. ISBN: 0-471-49926-9

4. Miwa K, Shimizu K, Fukui H (1989) Electrode substrate for fuel cell and process for producing same. US Patent 4,851,304, 25 July 1989
5. Breault R (2006) Electrode substrate for electrochemical cell from Carbon and cross-linkable resin fibers. PCT No: PCT/US2006/041494. 23 Oct 2006
6. Spearin W (1989) Process for forming a fuelcell matrix. European Patent 0,344,089 A1, 29 Nov 1989
7. Breault R (1977) Silicon carbide electrolyte retaining matrix for fuel cells. US Patent 4,017,664, 12 Apr 1977
8. Breault R, Mientek A, Sawyer R (1993) Minimized corrosion fuel cell device and method of making same. US Patent 5,270,132, 14 Dec 1993
9. Breault R, Fredley R, Scheffler G (1998) Corrosion resistant fuel cell assembly. US Patent 5,837,395, 17 Nov 1998
10. Reiser C, Landau M (1980) Method for reducing cell output voltage to permit low power operation. US Patent 4,202,933, 13 May 1980
11. Breault R, Harding R, Kemp F (1977) Method of fabricating a fuel cell electrode. US Patent 4,043,933, 23 Aug 1977
12. Goller G, Salonia J (1981) Dry floc method for making an electrochemical cell electrode. US Patent 4,287,232, 1 Sept 1981
13. Goller G, Salonia J (1982) Dry method for making an electrochemical cell electrode. US Patent 4,313,972, 2 Feb 1982
14. Dufner B (2008) Wettability ink, process and carbon composite articles made therewith. PCT No: PCT/US2007/007045, 25 Sept 2008
15. Breault R, Luoma W, Roche R (2008) Fuel cell separator plate assembly. US Patent 20,080,057,373, PCT No: PCT/US04/44007, 06 Mar 2008
16. Roche R (1993) Extruded fuel cell stack shunt current prevention arrangement. US Patent 5,178,968, 12 Jan 1993
17. Breault R (1983) Method for reducing electrolyte loss from an electrochemical cell. US Patent 4,414,291, 8 Nov 1983
18. Breault R, Martin R, Roche R, Kline R (1996) Cathode reactant flow field for a fuel cell stack. US Patent 5,558,955, 23 Sept 1996
19. Breault R (1980) Fuel cell electrolyte reservoir layer and method for making. US Patent 4,185,145, 22 Jan 1980
20. Breault R, Gorman M (1994) Laminated electrolyte reservoir plate. US Patent 5,366,825, 22 Nov 1994
21. Uemura T, Murakami S (1988) Method for producing a carbon sheet and a fuel cell separator. US Patent 4,737,421, 12 Apr 1988
22. Emanuelson R, Luoma W, Taylor W (1981) Separator plate for electrochemical cells. US Patent 4,301,222, 17 Nov 1981
23. Trocciola J, Schroll C, Elmore D (1975) Wet seal for liquid electrolyte fuel cells. US Patent 3,867,206, 18 Feb 1975
24. Schroll C (1974) Liquid electrolyte fuel cell with gas seal. US Patent 3,855,002, 17 Dec 1974
25. DeCasperis T, Roethlein R, Breault R (1981) Method of forming edge seals for fuel cell components. US Patent 4,269,642, 26 May 1981
26. Singelyn J, Gelting R, Mientek A (1988) Expanded high-temperature stable chemical resistant seal material. US Patent 4,774,154, 27 Sept 1988
27. Grevstad P (1976) Fuel cell cooling system with shunt current protection. US Patent 3,964,929, 22 June 1976
28. Breault R, Sawyer R, DeMarche T (1986) Cooling system for electrochemical fuel cell. US Patent 4,574,112, 4 Mar 1986
29. Breault R, Martin R, Roche R, Scheffler G, O'Brien J (2000) Coolant plate assembly for a fuel cell stack. US Patent 6,050,331, 18 Apr 2000
30. Grasso A, Martin R, Roche R (2000) Composite article. US Patent 6,039,823, 21 Mar 2000
31. Breault R, Fredley R (2007) Fuel cell with electrolyte condensation zone. US Patent Pub. No: US 2007/0224476 A1, 27 Sept 2007
32. Breault R, Rohrbach C (2007) Fuel cell assembly with operating temperatures for extended life. US Patent Pub. No: US 2007/0292725 A1, 20 Dec 2007
33. Breault R, Fredley R (2008) Fuel cell assembly having long life characteristics. US Patent Pub. No: US 2008/0118789 A1, 22 May 2008
34. Congdon J, English J (1986) Method and apparatus for adding electrolyte to a fuel cell stack. US Patent 4,596,749, 24 June 1986
35. Grevstad P (1986) Process for adding electrolyte to a fuel cell stack. US Patent 4,612,262, 16 Sept 1986
36. Ferro J (2009) PAFC history and successes. In: MCFC & PAFC R&D workshop, 2009 Fuel cell seminar, 16 Nov 2009
37. Facility type: Supermarkets and grocery stores. In: Energy star<sup>®</sup> building manual, chap 11. [http://www.energystar.gov/ia/business/EPA\\_BUM\\_CH11\\_Supermarkets.pdf](http://www.energystar.gov/ia/business/EPA_BUM_CH11_Supermarkets.pdf)
38. Supermarkets: An overview of energy use and energy efficiency opportunities, Energy star<sup>®</sup>
39. Jarvi T, Kanuri S (2010) Progress in phosphoric acid fuel cells. FC Expo, Tokyo, 5 Mar 2010
40. Gang X et al (1993) Electrolyte additives for phosphoric acid fuel cells. J Electrochem Soc 140(4):896–902
41. Jarvi T (2009) Fuel cells for combined heat and power applications. University of Pennsylvania, 26 Sept 2009
42. Norby T (1999) Solid-state protonic conductors: principles, properties, progress and prospects. Solid State Ionics 125:1–11
43. Grevstad P, Gelting R (1976) Fuel cell cooling system using a non-dielectric coolant. US Patent 3,969,145, 13 July 1976
44. Kanuri S (2009) PAFC cost challenges. In: MCFC & PAFC R&D workshop, 2009 Fuel cell seminar, 16 Nov 2009
45. Kanuri S UTC Power response to DOE Request for Information DE-FOA-0000225
46. Goller G, Salonia J, Petraglia V (1979) Dry mix method for making an electrochemical cell electrode. US Patent 4,175,055, 20 Nov 1979
47. <http://www.fuelcells.org/InternationalH2-FCpolicyfunding.pdf>
48. Kanuri S, Motupally S (2011) Engineering and application of phosphoric acid fuel cell system. ICEPAG, Costa Mesa, CA

## Photo-catalytic Hydrogen Production

JIEFANG ZHU

Department of Materials Chemistry, Ångström Laboratory, Uppsala University, Uppsala, Sweden

### Article Outline

Glossary

Definition of the Subject

Introduction

Principle of Photocatalytic H<sub>2</sub> Production

Evaluation of Photocatalytic H<sub>2</sub> Production

Experimental Setup for Photocatalytic H<sub>2</sub> Production

Photocatalysts for H<sub>2</sub> Production

Development and Modification of H<sub>2</sub> Production

Photocatalysts

Future Directions

Bibliography

### Glossary

**Band gap** In solid-state physics, a band gap, also called an energy gap, is an energy range in a solid where no electron states can exist. In graphs of the electronic band structure of solids, the band gap generally refers to the energy difference (in electron volts) between the top of the valence band and the bottom of the conduction band in insulators and semiconductors. This is equivalent to the energy required to free an outer shell electron from its orbit about the nucleus to become a mobile charge carrier, able to move freely within the solid material.

**Conduction band** In the solid-state physics field of semiconductors and insulators, the conduction band is the range of electron energies, higher than that of the valence band, sufficient to free an electron from binding with its individual atom and allow it to move freely within the atomic lattice of the material. Electrons within the conduction band are mobile charge carriers in solids, responsible for conduction of electric currents in metals and other good electrical conductors.

**Dopant** A dopant, also called a doping agent, is a trace impurity element that is inserted into a substance

(in very low concentrations) in order to alter the electrical properties or the optical properties of the substance. In the case of crystalline substances, the atoms of the dopant very commonly take the place of elements that were in the crystal lattice of the material.

**Doping** In semiconductor production, doping is the process of intentionally introducing impurities into an extremely pure (also referred to as intrinsic) semiconductor to change its electrical properties.

**Dye sensitization** The process in which the dye absorbs light to yield an excited state, which in turn transfers an electron (or energy) onto the semiconductor.

**Hydrogen** The chemical element with atomic number 1. It is represented by the symbol H. At standard temperature and pressure, hydrogen is a colorless, odorless, nonmetallic, tasteless, highly combustible diatomic gas with the molecular formula H<sub>2</sub>.

**Hydrogen production** The industrial method for generating hydrogen.

**Photocatalysis** In chemistry, photocatalysis is the acceleration of a photoreaction in the presence of a catalyst (called photocatalyst). In catalyzed photolysis, light is absorbed by an adsorbed substrate (called photocatalyst). In photogenerated catalysis, the photocatalytic activity depends on the ability of the catalyst (called photocatalyst) to create electron-hole pairs, leading to secondary reactions.

**Photocatalyst** A substance that is able to promote, by absorption of light quanta, chemical transformations of the reaction participants, repeatedly participating in intermediate chemical interactions and regenerating its chemical composition after each cycle of such interactions.

**Photocatalytic water splitting** The production of hydrogen (H<sub>2</sub>) and oxygen (O<sub>2</sub>) from water by directly utilizing the energy from light. Photocatalytic water splitting has the advantage of the simplicity of using powder or film photocatalysts in solution and (sun)light to produce H<sub>2</sub> and O<sub>2</sub> from water.

**Semiconductor** A material that has an electrical conductivity due to flowing electrons (as opposed to ionic conductivity) which is intermediate in magnitude between that of a conductor and an

insulator. This means roughly in the range  $10^{-8}$ – $10^3$  S/cm. In semiconductors, current is often schematized as being carried either by the flow of electrons or by the flow of positively charged “holes” in the electron structure of the material. Actually, however, in both cases only electron movements are involved.

**Valence band** In solids, the valence band is the highest range of electron energies where electrons are normally present at absolute zero temperature. The valence electrons are bound to individual atoms, as opposed to conduction electrons, which can move freely within the atomic lattice of the material. On a graph of the electronic band structure of a material, the valence band is located below the conduction band, separated from it in insulators and semiconductors by a band gap. In metals, the conduction band has no energy gap separating it from the valence band.

### Definition of the Subject

Increasing environmental concerns from using non-sustainable fossil fuels and a growing energy demand are forcing human beings to pursue clean and sustainable sources of energy. Hydrogen exists as a light, diatomic gas with high energy content by weight, but small energy content by volume. Hydrogen can react cleanly with oxygen in a highly exothermic reaction, with pure water as the only product. The energy stored in the chemical bond of hydrogen can be released by simply burning it in a combustion engine, or more efficiently by oxidizing it in a fuel cell. Based on the above properties, hydrogen is often considered an attractive alternative to the current, fossil fuel-based energy carrier. However, the currently most widely used methods to produce hydrogen are based on the conversion of fossil fuel resources and thus deviate completely from the environmentally friendly intention of using hydrogen.

Photocatalysis is a technique to convert light (ideally sunlight) energy to chemical energy or electrical power. Photocatalytic  $H_2$  production is a technique utilizing the energy from (sun)light to produce  $H_2$  from water (or other hydrogen resources). Photocatalytic  $H_2$  production from water has been accepted as one of the most promising ways to realize

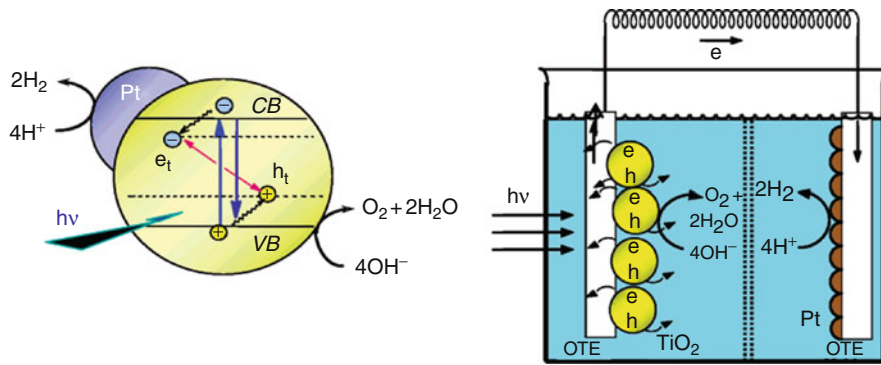
a hydrogen economy for three reasons: (1) this technology is based on photon (or solar) energy, which is a clean, perpetual source of energy, and mainly water, which is a renewable resource; (2) it is an environmentally safe technology without undesirable by-products and pollutants; and (3) the photochemical conversion of solar energy into a storable form of energy, i.e., hydrogen, allows one to deal with the intermittent character and seasonal variation of the solar influx [1].

Photocatalytic water splitting has the advantage of simply using powder or film photocatalysts, and in this entry, it excludes the contents on photoelectrochemical (PEC) water splitting, which is described in another entry of this encyclopedia. The redox mechanism of photocatalytic  $H_2$  production is similar to that of PEC  $H_2$  generation. The main difference between the two approaches lies in the location of the redox reactions, as illustrated in Fig. 1. In the photocatalytic process, both oxidation and reduction reactions occur on the surface of a photocatalyst (powder or film), and as a result, a mixture of  $H_2$  and  $O_2$  is evolved together (the left-hand side of Fig. 1). In the photoelectrochemical process, oxidation and reduction take place at spatially separated photoanode and cathode, respectively, resulting in  $H_2$  and  $O_2$  being evolved separately (the right-hand side of Fig. 1). It should be noted that the efficiency of photocatalysts is normally lower than that of photoanodes in  $H_2$  generation, since hydrogen and oxygen have a tendency to react back to water if they are evolved at the same location. However, compared to photoanodes, photocatalysts do not need a conductive substrate for charge collection, so that a much broader selection of synthetic methods, such as solid-state high temperature synthesis, can be adopted. This allows photocatalysts to be prepared with relative ease and at a competitive cost. In a photocatalytic system, cocatalysts can be easily introduced by firing and mixing. Furthermore, research on photocatalysts may provide a convenient screening approach for the selection of suitable photoanodes.

### Introduction

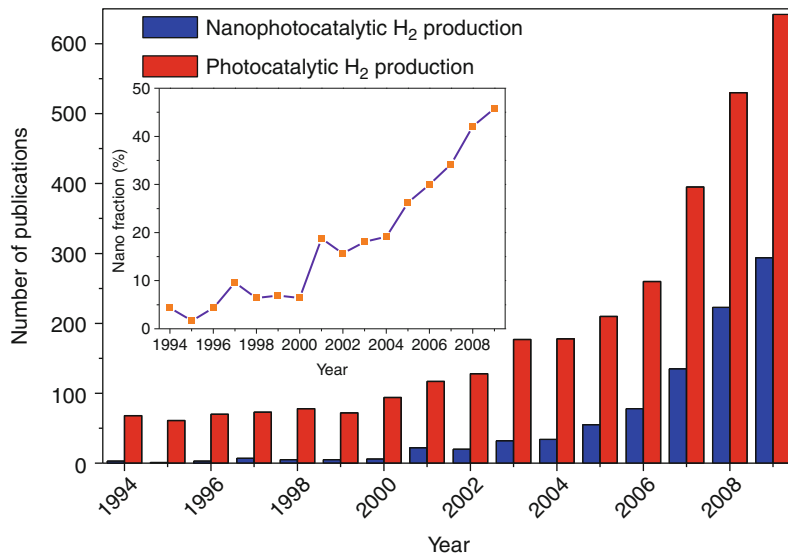
Photocatalytic  $H_2$  production has been developing for 40 years. The increasing number of scientific publications constitutes clear bibliographical evidence for the





**Photo-catalytic Hydrogen Production. Figure 1**

Photocatalytic splitting of water following the band gap excitation of the semiconductor nanoparticle (*left-hand side*) and a photoelectrolysis cell based on a nanostructured semiconductor film electrode (*right-hand side*) (Reprinted with permission from [2]. Copyright 2007 American Chemical Society)



**Photo-catalytic Hydrogen Production. Figure 2**

The number of publications on (nano)photocatalytic H<sub>2</sub> production sorted by year. The inset shows the fraction of publications on photocatalytic H<sub>2</sub> production that deal with nano-aspects (Data were collected from the “Web of Science,” and entries until October 14, 2010, have been considered)

significance of this hot field, as shown in Fig. 2. Recently, nanoscience and nanotechnology opened a new vista in this field. Since 2004, the number of publications on nanophotocatalytic H<sub>2</sub> production has increased by a factor of about 1.5 times every year. Many papers recently studied the impact of different nanostructures and nanomaterials on the performance of photocatalysts, since it was found that the total

energy conversion efficiency is largely determined by nanoscale properties of photocatalysts. Although progress has been made in many fields concerning photocatalytic H<sub>2</sub> production during these years, such as photocatalytic reactor design, product separation and detection, and light harvesting, the key factor determining its practical application is the materials, i.e., photocatalysts.

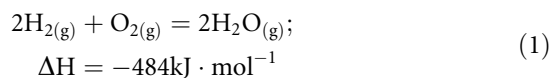
In 1972, the initial work of Fujishima and Honda indicated that PEC water splitting could be performed by using a TiO<sub>2</sub> photoanode [3]. The setup in that work is similar to that shown on the right-hand side of Fig. 1. Photogenerated electrons and holes are formed in TiO<sub>2</sub> when it is irradiated by UV light. The photogenerated electrons flow to a Pt counter electrode to reduce water to H<sub>2</sub>, while water is oxidized to O<sub>2</sub> by photogenerated holes on the TiO<sub>2</sub> side. Following this exciting finding, H<sub>2</sub> production by semiconductor photocatalysis and electrolysis has been attracting more and more attention. During the 1970s and the first half of the 1980s, the photocatalysts used for H<sub>2</sub> production from water splitting mainly focused on TiO<sub>2</sub>, SrTiO<sub>3</sub> [4, 5], and ZnO. Compared to TiO<sub>2</sub>, SrTiO<sub>3</sub> can split water without an external bias due to its higher conduction band level. In the middle of the 1980s, Pt/CdS [6–8] and ZnS [9] were identified as highly active photocatalysts for H<sub>2</sub> evolution under visible-light irradiation in the presence of sacrificial reagents. In the second half of the 1980s, new photocatalysts, such as K<sub>4</sub>Nb<sub>6</sub>O<sub>17</sub> [10–13], Na<sub>2</sub>Ti<sub>3</sub>O<sub>7</sub> [14], K<sub>2</sub>Ti<sub>2</sub>O<sub>5</sub> [14], and K<sub>2</sub>Ti<sub>4</sub>O<sub>9</sub> [14], were used for water splitting. Many tantalate [15–19], tungstate [20, 21], and molybdate [20] photocatalysts have shown high activity for H<sub>2</sub> production from an aqueous solution containing a sacrificial reagent since the second half of the 1990s. After coming into the new century, the database of photocatalysts for water splitting has become more plentiful. Many oxide photocatalysts consisting of the metal cations Ga<sup>3+</sup>, In<sup>3+</sup>, Ge<sup>4+</sup>, Sn<sup>4+</sup>, and Sb<sup>5+</sup>, with d<sup>10</sup> configuration (i.e., metal cations having fully filled (with ten electrons) outermost d orbitals), and assisted with RuO<sub>2</sub> as a cocatalyst have recently been reported [22–26]. Non-oxide Ge<sub>3</sub>N<sub>4</sub> with a RuO<sub>2</sub> cocatalyst was also found to be a photocatalyst for H<sub>2</sub> production [27]. Oxynitrides [28–31] and oxysulfides [32, 33] have been widely developed in research led by K. Domen for H<sub>2</sub> and O<sub>2</sub> evolution under visible-light irradiation in the presence of sacrificial reagents. Cr-Rh oxide/GaN:ZnO is a solid solution of GaN and ZnO, which is also active for overall water splitting [34]. The solid solution of ZnO and ZnGeN<sub>2</sub>, (Zn<sub>1+x</sub>Ge)(N<sub>2</sub>O<sub>x</sub>) was recently discovered as another active d<sup>10</sup> metal oxynitride photocatalyst for pure water splitting under visible light [35]. A Z-scheme photocatalytic water splitting

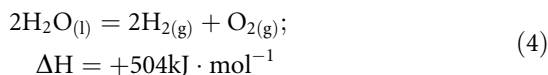
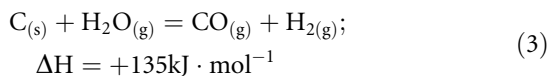
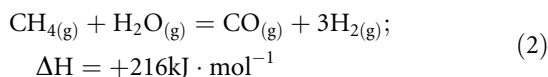
system, which involved two-step photoexcitation under visible-light irradiation, was recently developed by mimicking the natural photosynthesis of green plants [36, 37].

### Principle of Photocatalytic H<sub>2</sub> Production

As mentioned above, H<sub>2</sub> is an energy carrier, and the combustion of it produces only water and traces other pollutants with a large amount of heat releasing (Eq. 1). Moreover, H<sub>2</sub> is also an important raw material in chemical industries, e.g., industrial ammonia synthesis. However, a large fraction of hydrogen production is currently based on fossil fuel resources. The processes involved in the conversion of fossil fuels, such as the reforming of natural gas (Eq. 2), coal gasification (Eq. 3)/liquefaction, and coal electrolysis, require large amounts of energy, either in the form of heat or electricity, and all these processes are accompanied with the release of vast amounts of carbon dioxide. In order to actualize the hydrogen economy, there is a demand for other carbon-neutral feedstocks from which hydrogen can be produced with the energy provided by a sustainable source. Alternative feedstocks under consideration include wood, other biomass, organic waste, and water, where the former options are not necessarily carbon neutral, however. The latter option constitutes an entirely sustainable energy system, where hydrogen is produced from water (Eq. 4) and later on recovered back into water in a combustion engine or a fuel cell. This entry therefore focuses on hydrogen production from water.

In principle, water can be split into H<sub>2</sub> and O<sub>2</sub> by several different pathways and utilizing various energy sources. The commonly used method is to dissociate water in an electrolysis cell. However, electrolysis is a two-step process, and the potential of achieving a highly efficient, simple, and cost-effective conversion is limited. A better idea is to directly split water in a single device into its component gases, without the production of electricity in advance. Photocatalysis can directly utilize photon energy to decompose water into H<sub>2</sub> and O<sub>2</sub>.

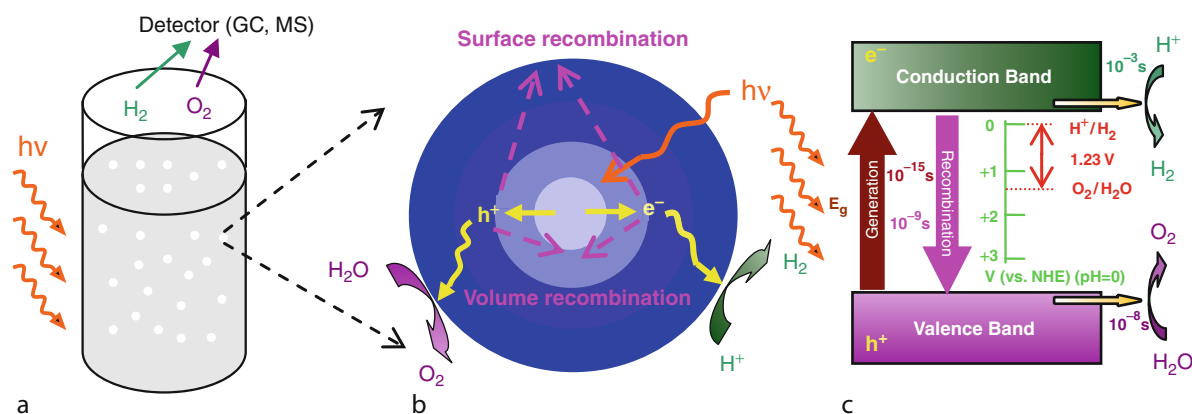




Photocatalytic  $\text{H}_2$  production (or water splitting) occurs on the surface of semiconductor materials, as shown in Fig. 3. Figure 3 shows three basic steps in photocatalytic  $\text{H}_2$  production. The first step is the absorption of photons. A semiconductor has a valence band (VB) and a conduction band (CB), which are separated from one another by a band gap ( $E_g$ ), as shown in Fig. 3c. In the ground state, all electrons exist in the VB. Under irradiation by photons with energy equivalent to or larger than  $E_g$ , some of the electrons are excited from the VB to the CB, leaving empty states, so-called holes, in the VB. The second step is the charge separation and migration, as shown in Fig. 3b. The photogenerated electrons and holes from the first step may recombine in the bulk or on the surface of the semiconductor on a time scale which is slower than the time required for their generation (Fig. 3b, c). Some electrons and holes that travel to the

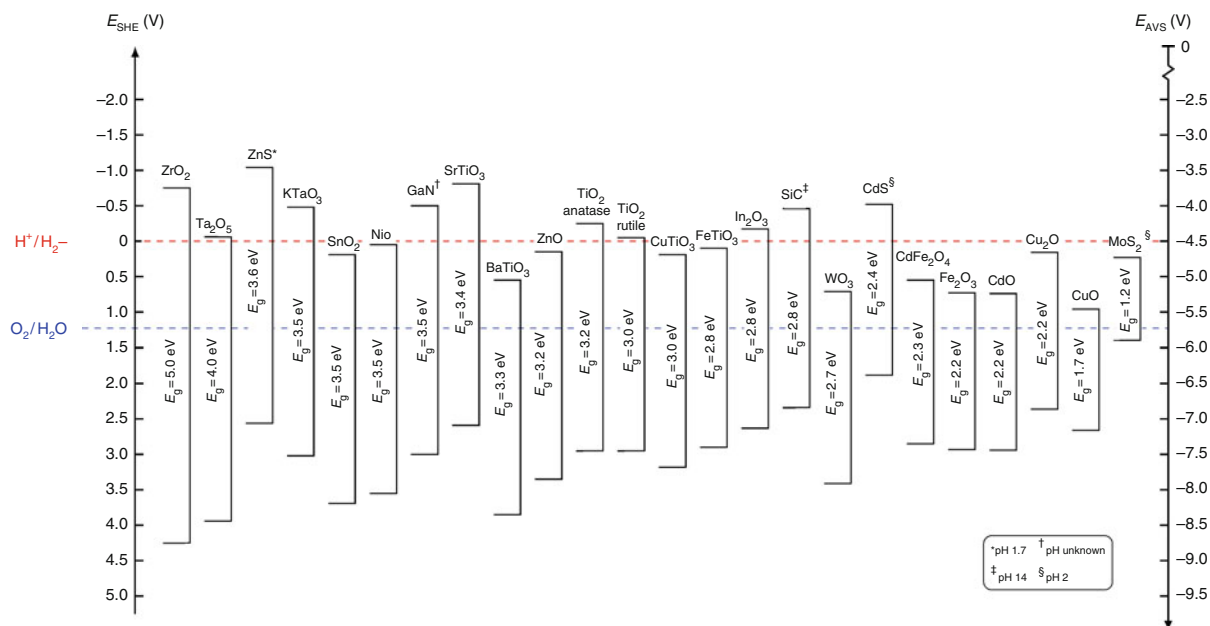
surface of the semiconductor without recombination can cause reduction ( $\text{H}_2$  formation) and oxidation ( $\text{O}_2$  formation) reactions, respectively, in the last step called surface chemical reactions. The important parameters in photocatalysts are the width of the band gap and the levels of the conduction and valence bands. From Fig. 3c, the theoretical minimum band gap (or photon energy) for water splitting is 1.23 eV, corresponding to a wavelength of 1,008 nm, according to the formula  $E_g$  (eV) equal to  $1,240/\lambda$  (nm).

However, if taking into account thermodynamic losses at various steps in the photocatalytic process, overpotential is necessary to ensure a reasonable reaction rate, and effective photocatalysts exhibit band gaps larger than 2 eV, corresponding to a wavelength below 620 nm. If the sunlight is to be used for water splitting, visible-light-responsive photocatalysts should have band gaps between 2 and 3.1 eV, since the intensity of sunlight is small in the UV region, i.e., for wavelengths below 400 nm (3.1 eV). Apart from the band gap requirement, for  $\text{H}_2$  production to take place, the CB bottom edge should be more negative than the reduction potential of  $\text{H}^+/\text{H}_2$  ( $E_{\text{H}^+/\text{H}_2} = 0\text{V}$  vs NHE at pH 0), while the VB top edge should be more positive than the oxidation potential of  $\text{O}_2/\text{H}_2\text{O}$  ( $E_{\text{O}_2/\text{H}_2\text{O}} = 1.23\text{V}$  vs NHE at pH 0) for  $\text{O}_2$  evolution from water to occur, as shown in Fig. 3c. Figure 4 lists the band edge positions for



**Photo-catalytic Hydrogen Production. Figure 3**

(a) A suspension of photocatalyst powders in water under irradiation. (b) The processes involved in water splitting by a photocatalyst particle are photon absorption, electron-hole generation (and recombination), charge transport, and oxidation/reduction reactions on the semiconductor surface. (c) The principle and energy diagram for photocatalytic water splitting on a semiconductor (Reprinted from [1], Copyright (2009), with permission from Elsevier)



**Photo-catalytic Hydrogen Production. Figure 4**

Valence and conduction band edge positions of various semiconductors with respect to the standard hydrogen electrode (SHE) scale and the vacuum reference energy scale  $E_{AVS}$ . All values were reported being tested at pH 1, unless otherwise noted

commonly used semiconductors. It can be seen that there are not many semiconductors meeting the requirements of band gap and band levels for ideal photocatalysts for water splitting. Besides thermodynamic requirements, an ideal photocatalyst should also have kinetic advantages, e.g., high photocatalytic activity for H<sub>2</sub> and O<sub>2</sub> production. There are a few non-oxide semiconductors that meet requirements mentioned above. However, they are unstable and/or easily corroded (e.g., CdS). It has proven difficult to find a simple and cost-effective photocatalyst meeting all the requirements [1].

### Evaluation of Photocatalytic H<sub>2</sub> Production

There is no perfect standard for comparing H<sub>2</sub> productivity of different photocatalysts in different photocatalytic systems. Therefore, researchers normally compare their homemade photocatalysts in their setups with some references, e.g., commercial TiO<sub>2</sub> photocatalysts, which are well known for their high activity, like P25 (a product of Evonik Degussa GmbH), Hombikat UV-100 (a product of Sachtleben

Chemie GmbH), and ST-01 (a product of Ishihara Sangyo Kaisha, Ltd.). However, there are some terms that are commonly used to describe photocatalytic activity, and these are discussed below.

### Turnover Quantities

The quantitative measurements of photocatalytic activity of a solid photocatalyst are normally derived from process kinetics to express a rate referenced to the number of photocatalytic sites to infer how many times a catalytic cycle turnovers. Traditionally, these parameters are the turnover number (TON), turnover rate (TOR), and turnover frequency (TOF). In H<sub>2</sub> production, turnover rate (TOR) is considered as the number of H<sub>2</sub> molecules produced per active site per unit time (units: molecules site<sup>-1</sup> time<sup>-1</sup>) (Eq. 5); turnover frequency (TOF) as the number of H<sub>2</sub> molecules produced per unit time (units: molecules time<sup>-1</sup>) (Eq. 6); and turnover number (TON) as a quantity that describes how many times a H<sub>2</sub> evolution reaction or process turnovers at active sites integrated over time (units: molecules site<sup>-1</sup>) (Eq. 7).

$$\text{TOR} = \frac{\text{Number of H}_2 \text{ molecules produced}}{\text{Number of activesites} \cdot \text{Reaction time}} \quad (5)$$

$$\text{TOF} = \frac{\text{Number of H}_2 \text{ molecules produced}}{\text{Reaction time}} \quad (6)$$

$$\text{TON} = \frac{\text{Number of H}_2 \text{ molecules produced}}{\text{Number of activesites}} \quad (7)$$

Using these turnover quantities, it can be judged whether a given process is truly photocatalytic, as TON of a real photocatalytic reaction is much higher than 1. In addition, these turnover quantities are also supposed to be useful in assessing new materials as photocatalysts, and reproducible or comparable across various laboratories. Normally, it is difficult to determine the number of active sites for a photocatalyst, and the active sites and non-active sites can switch during photocatalytic process. Therefore, the number of atoms in a photocatalyst or on the surface of a photocatalyst, instead of the number of active sites, is often employed in Eqs. 5 and 7. However, the irradiated surface area is not equal to the total surface area of the photocatalyst, and the active sites can only lie in the irradiated surface area. Thus, the practical determination of turnover quantities remains very complex, and cannot be accurate.

### Quantum Yield and Photonic Efficiency

Quantum yield ( $\Phi$ ) is defined as the number of defined events that occur per photon absorbed by the system, or as the amount (mol) of reactant consumed or product ( $\text{H}_2$ ) formed per amount of photons (mol or Einstein) absorbed (Eq. 8). The definition of quantum yield makes it difficult to describe photocatalytic efficiency in real heterogeneous media, particularly for complex reactor geometries. In suspension systems, the sum of reflection, scattering, and transmission should be measured precisely, in order to determine the amount of photons absorbed by the photocatalyst.

$$\Phi = \frac{\text{Number of H}_2 \text{ molecules produced}}{\text{Number of photons absorbed by photocatalyst}} \quad (8)$$

A simple alternative method of comparing process efficiencies for equal absorption of photons has been proposed by N. Serpone for heterogeneous photocatalysis: (relative) photonic efficiency [38].

Photonic efficiency ( $\xi$ ) describes the number (or mols) of reactant molecules transformed or product ( $\text{H}_2$ ) molecules formed divided by the number or Einsteins of photons at a given wavelength incident on the reactor cell (Eq. 9). Alternatively, the photonic efficiency can also be described by the ratio of the initial rate of the event to the rate of incident photons reaching the reactor as obtained by actinometry (Eq. 10).

$$\xi = \frac{\text{Number of H}_2 \text{ molecules produced}}{\text{Number of incident photons on reactor}} \quad (9)$$

$$\xi = \frac{\text{The rate of H}_2 \text{ production}}{\text{The rate of photons impinging on reactor}} \quad (10)$$

To avoid unnecessary errors and the effects from reactor geometry and light source, together with the properties (e.g., size, surface area) of the photocatalyst material used, another kind of efficiency has been defined so that it could be used to compare experiments within the same laboratory or even with other laboratories, and it would be reactor independent: the relative photonic efficiency ( $\xi r$ ). It is related to an acceptable standard process, a standard photocatalyst material, or a standard “secondary actinometer” in photocatalytic processes (Eq. 11). For example, the determination of the total incident photon in the wavelength regions by chemical actinometry ferrioxalate has been performed in the same reactor, in order to avoid the corrections for any influence of light reflection, beam position, and reactor geometry. In the experimental description of a relative photonic efficiency, reactor geometry, light source, and photocatalyst properties should be constant in assessing  $\xi r$ . To be really useful and comparable,  $\xi r$  values should not depend on light irradiance and reactor geometry, or even on other parameters such as pH, photocatalyst loading, substrate concentration, and temperature [38], which is not easy in practice.

$$\xi r = \frac{\text{The rate of H}_2 \text{ production}}{\text{The rate of standard process under identical conditions}} \quad (11)$$

### Experimental Setup for Photocatalytic $\text{H}_2$ Production

There are several kinds of setups for photocatalytic  $\text{H}_2$  production and/or water splitting. Generally speaking,

these setups consist of four parts. The first part is a light source. A high-pressure mercury lamp is often used with quartz apparatuses for broad band gap semiconductors, since UV-light irradiation is needed. A xenon lamp is used for visible-light irradiation, when a cut-off filter is employed to avoid infrared heating. A solar simulator (AM1.5) with its radiance of  $100 \text{ mW}\cdot\text{cm}^{-2}$  is ideal for solar hydrogen production study. Alternatively, a xenon lamp and several filters can be assembled for a solar simulator. In this case, the emission spectrum should be measured to make sure of its identity to solar spectrum, and the intensity should be calibrated by a thermopile or Si photodiode. The second part is a reaction cell. There are different shapes and sizes of reactors for photocatalytic  $\text{H}_2$  production and/or water splitting. They should be transparent enough for light irradiation, and they are connected to the third part, a gas line. Vacuum pump, pressure gauge, and carrier gas are normally connected to the gas line. The last part is gas detection and/or collection part, which draws in the produced gases ( $\text{H}_2$  and  $\text{O}_2$ ) from gas line. Gas chromatography and mass spectrometry are employed for microanalysis, while volumetric measurement is suitable for a large amount of gas evolution. The whole system (including reactor, gas line, gas detector, and collector) should stay gas tight during operation.

## Photocatalysts for $\text{H}_2$ Production

Most photocatalysts are composed of both metal and nonmetal ions. Metal cations show their highest oxidative states with  $d^0$  (red area in Fig. 5) or  $d^{10}$  (green area) electronic configuration, while O, S, and N (blue area) exist as their most negative states. The conduction band bottom consists of the d and sp orbitals of the metal cations, while the valence band top in metal oxides is composed of O 2p orbitals, which is normally located at ca. +3 V (vs SHE) or higher. The valence bands of metal oxysulfides and oxynitrides are formed by S 3p and O 2p, and N 2p and O 2p, respectively. In some compounds, alkali (Li, Na, K, Rb, and Cs), alkaline earth (Mg, Ca, Sr, and Ba), and transition metal (Y, La, and Gd) ions can constitute the crystal structures of these compounds, rather than make any contribution to the energy structures. In this section, the commonly used materials for photocatalytic  $\text{H}_2$  production will be introduced.

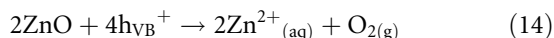
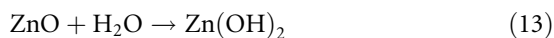
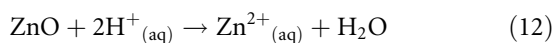
$\text{TiO}_2$  has been one of the most important photocatalysts for  $\text{H}_2$  production, due to its availability, low price, nontoxicity, high photoactivity, and stability. The biggest disadvantage of  $\text{TiO}_2$  is the low utilization of visible light, as a consequence of its wide band gaps, which are 3.2 eV (corresponding to an absorption edge of 380 nm) and 3.0 eV (400 nm) for

H																	He
Li	Be											B	C	N	O	F	Ne
Na	Mg											Al	Si	P	S	Cl	Ar
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba	Ln	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
			d <sup>0</sup> configuration					d <sup>10</sup> configuration									
La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu			

Photo-catalytic Hydrogen Production. Figure 5

Elements constructing photocatalysts (Reprinted from [1], Copyright (2009). With permission from Elsevier)

anatase and rutile, respectively. At any rate, TiO<sub>2</sub> is a good reference and starting material for research. ZnO has the similar advantages (including availability, low price, nontoxicity, high photoactivity, and enough potential for both H<sub>2</sub> and O<sub>2</sub> evolution) with TiO<sub>2</sub> and single-crystalline ZnO even has fast electron mobility. However, it is unstable in strong acids (Eq. 12) and alkalis (Eq. 13), and after long-time irradiation, ZnO suffers from photo-corrosion, due to the oxidation of O<sup>2-</sup> in ZnO by photogenerated holes (Eq. 14). WO<sub>3</sub> has a relatively narrow band gap of 2.7 eV, but enables it to utilize part of the visible light. WO<sub>3</sub> is a successful photocatalyst for O<sub>2</sub> evolution from water, while its conduction band is not negative enough to reduce H<sup>+</sup> to H<sub>2</sub>. In practice, WO<sub>3</sub> is applied to water splitting by coupling to another semiconductor or by doping. α-Fe<sub>2</sub>O<sub>3</sub> has several advantages of narrow band gap of 2.2 eV, good photo stability, chemical inertness, and low cost. However, it has the same problem as WO<sub>3</sub> with respect to its positive conduction band level. Furthermore, it suffers from fast e<sup>-</sup>-h<sup>+</sup> recombination and poor charge transportation.



When TiO<sub>2</sub> is fused with other metal oxides (SrO, BaO, Ln<sub>2</sub>O<sub>3</sub> (Ln = lanthanide)), metal titanates with perovskite structure are formed. Perovskites have the general formula ABX<sub>3</sub>, and several hundred oxides own this structure. Among them, SrTiO<sub>3</sub> and BaTiO<sub>3</sub> both with a band gap of 3.3 eV have been widely studied as semiconductors for photocatalytic water splitting. Alkaline metal hexatitanates (M<sub>2</sub>Ti<sub>6</sub>O<sub>13</sub>; M = Na, K, Rb) are normally used in powder form in suspensions, together with a cocatalyst. There are more complex perovskites containing two different cations, and many of these have a layered structure. Two main classes of such oxides, which have been studied in water splitting, are the Dion–Jacobson series (AM<sub>n-1</sub>B<sub>n</sub>O<sub>3n+1</sub>, e.g., KCa<sub>2</sub>Ti<sub>3</sub>O<sub>10</sub>), and the Ruddlesden–Popper series (A<sub>2</sub>M<sub>n-1</sub>B<sub>n</sub>O<sub>3n+1</sub>, e.g., K<sub>2</sub>La<sub>2</sub>Ti<sub>3</sub>O<sub>10</sub>). Another type of layered perovskites has the generic composition A<sub>n</sub>B<sub>n</sub>O<sub>3n+2</sub> (n = 4, 5; A = Ca, Sr, La; B = Nb, Ti). Among them, La<sub>2</sub>Ti<sub>2</sub>O<sub>7</sub> (i.e., La<sub>4</sub>Ti<sub>4</sub>O<sub>14</sub>) and

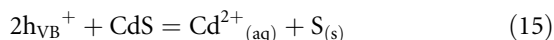
La<sub>4</sub>CaTi<sub>5</sub>O<sub>17</sub> are representative titanates. The band gaps of La<sub>2</sub>Ti<sub>2</sub>O<sub>7</sub> and La<sub>4</sub>CaTi<sub>5</sub>O<sub>17</sub> are 3.2 and 3.8 eV, and when loaded with nickel they showed a quantum yield of 12% (<360 nm) and 20% (<320 nm) for water splitting, respectively [39].

Tantalates and niobate oxides with corner-sharing octahedral MO<sub>6</sub> (M = Ta, Nb) structures have shown high photocatalytic activity for the cleavage of water, since the photogenerated electron–hole pairs can easily migrate and separate through the corner-shared MO<sub>6</sub> units. Tantalates, MTaO<sub>3</sub> (M = Li, Na, K) were reported as effective photocatalysts for water splitting under UV irradiation. These oxides own perovskite structure, and their band gaps are 4.7 eV (Li), 4.0 eV (Na), and 3.7 eV (K), respectively [40]. Some layered oxides (e.g., K<sub>4</sub>Nb<sub>6</sub>O<sub>17</sub>) have two kinds of interlayers, which exist alternately. H<sub>2</sub> is evolved from one interlayer, where cocatalysts are selectively introduced. In K<sub>4</sub>Nb<sub>6</sub>O<sub>17</sub> case, H<sub>2</sub> and O<sub>2</sub> evolutions are separated by the photocatalytic niobate sheet, which will be described later.

Generally speaking, oxides containing d<sup>0</sup> transition metal cations, like Ti<sup>4+</sup>, Nb<sup>5+</sup>, or Ta<sup>5+</sup>, have wide band gaps (>3.0 eV). Thus, these materials can only be excited under UV irradiation. Interestingly, K<sub>4</sub>Ce<sub>2</sub>M<sub>10</sub>O<sub>30</sub> (M = Ta, Nb) and their solid solutions K<sub>4</sub>Ce<sub>2</sub>Ta<sub>10-x</sub>Nb<sub>x</sub>O<sub>30</sub> (x = 0–10) were found to have band gaps of ca. 1.8–2.3 eV (corresponding to absorption edges of 540–690 nm), and they performed well under visible-light irradiation. This may be ascribed to the hybrid construction of their valence bands [41].

Many metal sulfide photocatalysts are active under visible-light irradiation. The valence bands of metal sulfides consist of S 3p orbitals, which make the valence band potential more negative and thus narrows the band gap, compared to the valence bands composed of O 2p orbitals in corresponding metal oxides. CdS with wurtzite structure is the most studied metal sulfide photocatalyst. It has a narrow band gap (2.4 eV), which makes it absorb visible-light below a wavelength of 510 nm. Although the valence and conduction bands of CdS are perfectly suitable for O<sub>2</sub> and H<sub>2</sub> evolution, CdS is apt to be photo-corroded (Eq. 15), which is common for most metal sulfides. In order to avoid photo-corrosion, scavengers (electron donors, such as cysteines, EDTA, sulfide, or sulfite species) are often

used to consume photogenerated holes during the photocatalytic water splitting by CdS.



Enlightened by the N doping effect,  $\text{Ta}_3\text{N}_5$  was prepared by nitriding  $\text{Ta}_2\text{O}_5$  in an  $\text{NH}_3$  atmosphere. Reasonably, the band gap shrinks from  $\sim 4.0$  eV for  $\text{Ta}_2\text{O}_5$  to  $\sim 2.1$  eV for  $\text{Ta}_3\text{N}_5$ . The narrower band gap results from a higher-lying valence band derived from N 2p orbitals in  $\text{Ta}_3\text{N}_5$  other than O 2p orbitals in  $\text{Ta}_2\text{O}_5$ . This material photocatalytically produced  $\text{H}_2$  and  $\text{O}_2$  under visible irradiation ( $< 600$  nm) [42]. Nitrides with  $d^{10}$  electronic configuration, such as  $\text{Ge}_3\text{N}_4$  and  $\text{GaN}$ , also performed the cleavage of water under UV-light irradiation. In these  $d^{10}$  electronic configuration nitrides, the conduction band formed by broad hybridized sp orbitals makes it easy to transfer the photogenerated electrons to cocatalysts, e.g.,  $\text{RuO}_2$  [27, 43].

Oxynitrides and oxysulfides were recently designed to split water under visible-light irradiation. For oxynitrides and oxysulfides with  $d^0$  electronic configuration metal ions, such as  $\text{TaON}$  and  $\text{Sm}_2\text{Ti}_2\text{S}_2\text{O}_5$ , the valence band mainly consists of hybridized N 2p (or S 3p) and O 2p orbitals, while the conduction band is still composed of the empty d orbitals of the corresponding metal. In such compounds, photogenerated holes can move smoothly in the broad valence band, which benefits to the oxidation of water. Oxynitride photocatalysts consisting of  $d^0$  configuration metal cations such as  $\text{Ti}^{4+}$ ,  $\text{Nb}^{5+}$ , and  $\text{Ta}^{5+}$  are active for  $\text{H}_2$  or  $\text{O}_2$  evolution in the presence of sacrificial reagents.  $\text{TaON}$  prepared by partial nitridation of  $\text{Ta}_2\text{O}_5$  showed high activity under visible-light irradiation ( $420 \text{ nm} \leq \lambda \leq 500 \text{ nm}$ ). The band gap of  $\text{TaON}$  was estimated to be 2.5 eV. Photocatalytic  $\text{H}_2$  production by  $\text{TaON}$  was performed in an aqueous methanol solution and with a Ru cocatalyst [28, 30]. As mentioned above, compared to photocatalysts containing  $d^0$  metal ions, the photocatalysts composed of  $d^{10}$  metal ions (such as Ge and Ga) have the conduction band bottom made up of hybridized sp orbitals of  $d^{10}$  metal ions. These hybridized sp orbitals increase the mobility of photogenerated electrons in the conduction band and high photocatalytic activity for reduction of water. The solid solutions ( $\text{Ga}_{1-x}\text{Zn}_x$ )

( $\text{N}_{1-x}\text{O}_x$ ) [34] and ( $\text{Zn}_{1+x}\text{Ge}$ )( $\text{N}_2\text{O}_x$ ) [35] are two examples of these  $d^{10}$  metal oxynitrides.

P-type III–V semiconductors have several attractive features, such as a high charge carrier mobility, an ideal band gap, and high photoelectrochemical stability, which make them suitable photocatalytic materials for reducing water to  $\text{H}_2$ . P-InP photocathodes are capable of producing  $\text{H}_2$  from  $\text{HCl}$  or  $\text{HClO}_4$  electrolytes with high efficiency [44]. Photocathodes of p-GaInP<sub>2</sub> (a solid solution of GaP and InP) have also evolved  $\text{H}_2$  efficiently by using a GaAs p-n junction bias [45].

## Development and Modification of $\text{H}_2$ Production Photocatalysts

Whether photocatalytic  $\text{H}_2$  production can have practical application depends largely on the development of photocatalytic materials. That is why material research is the focus of this topic. Both the development of existing photocatalysts and the exploitation of new photocatalysts are necessary. The progress of related fields in materials science, chemistry, and physics also add new vigor to the field of photocatalytic  $\text{H}_2$  production.

### Nanosize Effect

With the development of advanced fabrication and characterization, nanosized photocatalysts have become the main object of material study. When a photocatalyst becomes smaller, it has a larger surface area, which provides more active sites for reactant adsorption and decomposition, and light harvest. In photocatalysts, photogenerated electrons and holes transfer to the surface to have effect. Small particles provide a short distance for the charge carrier transfer, largely avoiding the bulk recombination. When a particle is smaller than some critical size (normally in the nanosize range), the energy levels within the filled (valence band) and empty (conduction band) states become discrete, and simultaneously the band gap increases, compared to its bulk counterpart. This leads to a blue shift in the absorption spectra for nanosized particles. Different materials have different critical sizes. With the broadening of the band gap, electrons at the bottom edge of the conduction band and holes at the top edge of the valence band acquire more negative and positive potentials, respectively, which means that they have stronger redox powers in such nanoparticles.



## Porous Structures

Porous materials have high surface area and good adsorption ability, and can concentrate reactants around active sites. Selective photocatalysis can be achieved by adjusting the pore size. Recently, advanced fabrication techniques have been developed to prepare porous photocatalysts with high surface area and suitable pore size. It has been reported that hydrothermally synthesized TiO<sub>2</sub> nanoparticles without calcination had a large specific surface area (438 m<sup>2</sup>/g) and small crystallites (2.3 nm) dispersed among amorphous mesoporous domains, and exhibited much better photocatalytic activity for H<sub>2</sub> production compared with samples calcined at various temperatures, and also the commercial photocatalyst P25 [46].

A novel synthesis was carried out by using KCl electrolyte to control the electrostatic repulsive force between TiO<sub>2</sub> nanoparticles toward the formation of a mesoporous structure, which showed the highest photocatalytic activity for H<sub>2</sub> production, compared to nonporous colloidal TiO<sub>2</sub>, and commercial Degussa P25 and Hombikat UV-100 (HBK) samples [47]. Cocatalysts can also be easily deposited and dispersed onto these porous photocatalysts. The photocatalytic reduction of metal cations (M = Ni<sup>2+</sup>, Co<sup>2+</sup>, Cu<sup>2+</sup>, Cd<sup>2+</sup>, Zn<sup>2+</sup>, Fe<sup>2+</sup>, Ag<sup>+</sup>, Pb<sup>2+</sup>) on the surface of mesoporous TiO<sub>2</sub> (specific surface area 130–140 m<sup>2</sup>/g, pore diameter 5–9 nm, and anatase content 70–90%) resulted in the formation of nanostructured metal–semiconductor composites (TiO<sub>2</sub>/M). These metal–TiO<sub>2</sub> nanostructures showed a remarkable photocatalytic activity for hydrogen production from water–alcohol solutions, and the efficiency was 50–60% greater than that of the metal-containing nanocomposites based on Degussa P25. The anatase content and pore size proved to be the main parameters determining the photoreaction rate [48].

Porous materials are often used as supports for catalysts and photocatalysts. Highly dispersed and coordinated metal species on microporous zeolite and mesoporous silica materials have shown very high photocatalytic activity, and were referred to as “single-site photocatalysts.” Photocatalytic H<sub>2</sub> production by CdS has been improved by its porous supports, such as aluminum-substituted mesoporous silica

molecular sieve (Al-HMS) [49], microporous and mesoporous silicas [50], porous polyethylene terephthalate fibers (PET) [51], and ETS-4 zeolite [52].

## Low-Dimensional Nanostructure

One-dimensional (1D) nanostructures (nanowires, nanorods, nanotubes, and nanofibers) have shown their outstanding properties in photocatalytic H<sub>2</sub> production. Compared to aspheric nanoparticles with the same volume or weight, 1D nanomaterials have higher surface areas. They can normally provide fast charge transportation, especially for those single-crystalline 1D nanostructures. Anders et al. found that photoelectrodes with nanorods oriented perpendicular to the conductive substrate can shorten the transport distance for electrons to the back contact (electron collector) and avoid recombination losses at grain boundaries between nanoparticles, compared to photoelectrodes with nanoparticle deposits [53]. Disorder, order, and different orientations of the 1D units also affect the photoelectrochemical properties. In the same research [53], the reported photon-to-current efficiency is lower for the Fe<sub>2</sub>O<sub>3</sub> electrode with the nanorods parallel to the substrate, compared to nanorods perpendicular to the substrate. The electrons have a more straightforward pathway to the back contact with nanorods perpendicular to the substrate, which leads to the elimination of recombination losses at nanorod boundaries and higher incident photon-to-electron conversion efficiency (IPCE) values.

As mentioned above in the [section Photocatalysts for H<sub>2</sub> Production](#), the poor charge transportation (mainly due to low mobility of holes) of  $\alpha$ -Fe<sub>2</sub>O<sub>3</sub> largely limits its application in photoelectrochemical and photocatalytic H<sub>2</sub> production. One way to solve this problem is to use 1D  $\alpha$ -Fe<sub>2</sub>O<sub>3</sub> with high aspect ratios. It was demonstrated that the transportation distance for photogenerated holes to the  $\alpha$ -Fe<sub>2</sub>O<sub>3</sub>/solution interface was largely shortened in  $\alpha$ -Fe<sub>2</sub>O<sub>3</sub> nanowires [54]. The limitation from hole transportation can be reasonably overcome, if the  $\alpha$ -Fe<sub>2</sub>O<sub>3</sub> nanowire radius is shorter than the hole diffusion length. 1D nanostructured photocatalysts, e.g., TiO<sub>2</sub> nanotubes [55–58], nanocolumns [59], nanowires [60] and nanofibers [61], and CdS nanorods [62] and

nanowires [63], have been playing a very important role in photocatalytic  $\text{H}_2$  production.

Two-dimensional (2D) structured materials (nanosheets, nanoscrolls, and nanolayers) have attracted special interest in catalysis and photocatalysis. They can have a high surface area, expose a certain facet with high photocatalytic activity, and provide fast charge transfer. Pt/ $\text{TiO}_2$  nanosheets with exposed (001) facets showed high photocatalytic activity for  $\text{H}_2$  production [64]. High surface energy of (001) facets is effective for dissociative adsorption of reactant molecules, and water molecules can chemically dissociate on the (001) surface. So the authors believed that the exposed (001) facets contributed to the high photocatalytic activity in  $\text{H}_2$  production [64], which had been also suggested by Amano [65] and Lu et al. [66].

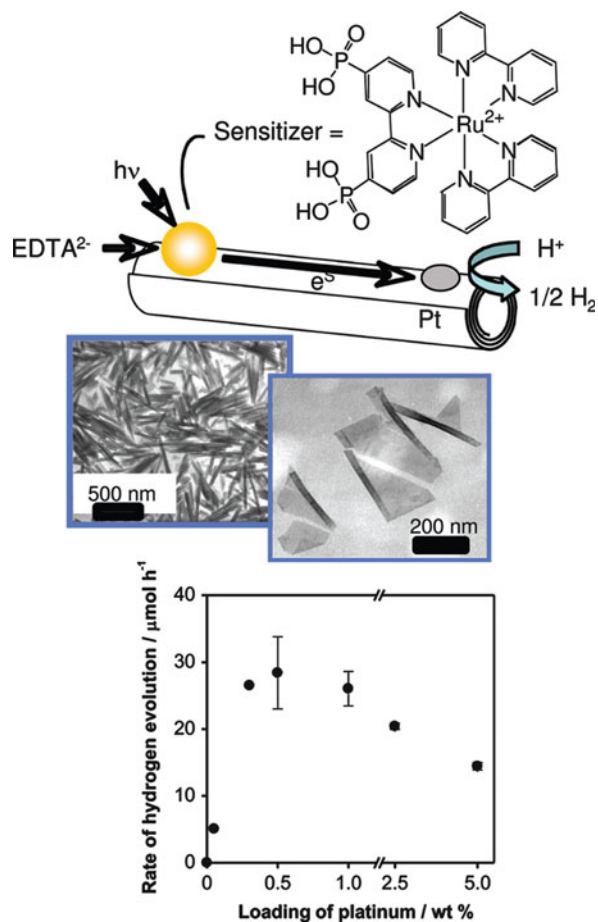
The following example shows the fast charge transfer of nanosheets and nanoscrolls. In a dye-sensitized semiconductor photocatalytic  $\text{H}_2$  production system, the semiconductor transfers electrons from the photoexcited dye to the hydrogen evolution catalyst (typically Pt or Rh). Niobate nanoscrolls and nanosheets acted as good electron transfer mediators between a phosphonated  $[\text{Ru}(\text{bpy})_3]^{2+}$  derivative and Pt, as shown in Fig. 6. An external quantum yield (incident photo-to-current yield) of 20–25% was observed for hydrogen evolution in this system [67].

Layered structures have been widely used in photocatalytic  $\text{H}_2$  production.  $\text{K}_4\text{Nb}_6\text{O}_{17}$  has a structure with two types of interlayers.  $\text{H}_2$  is produced from one interlayer, in which cocatalysts are introduced by ion exchange or interlayer reaction, while  $\text{O}_2$  is evolved in the other interlayer, as shown in Fig. 7. In this way, the sites for  $\text{H}_2$  and  $\text{O}_2$  evolution are separated by the photocatalytic niobate sheet [68].

### Dye-Sensitized Semiconductor

Some wide band semiconductors (such as  $\text{TiO}_2$ ,  $\text{SrTiO}_3$ , and  $\text{ZnO}$ ) show very high photocatalytic activity in  $\text{H}_2$  production under UV illumination, but cannot absorb visible light, which limits their application in solar energy conversion. One way to extend their light response to visible range is to use a dye, which can absorb visible light. This system is called a dye-sensitized semiconductor system. In this system, the

dye absorbs the visible light, becomes excited, and its excited state injects electrons into the semiconductor conduction band, on which  $\text{H}_2$  is produced usually in the presence of a metal cocatalyst. In order to regenerate dyes, electron donors, such as  $\text{I}_3^-/\text{I}^-$  pair and EDTA, are added into the solution to supply the dyes with electrons and sustain the reaction cycle. The excitation, electron injection, and dye regeneration can be expressed as in the following Eqs. 15–17. The steps

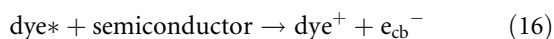


**Photo-catalytic Hydrogen Production.** Figure 6

(Top) Schematic representation of photoinduced electron transfer from a phosphonated  $[\text{Ru}(\text{bpy})_3]^{2+}$  sensitizer to Pt catalyst particles, mediated by  $\text{H}_4\text{Nb}_6\text{O}_{17}$  nanoscrolls. (Center) TEM images of individual nanosheets (right) and of nanoscrolls precipitated from a suspension of exfoliated  $\text{H}_4\text{Nb}_6\text{O}_{17}$  (left). (Bottom) The dependence of hydrogen evolution rate on Pt loading (Reprinted with permission from [67]. Copyright 2009 American Chemical Society)

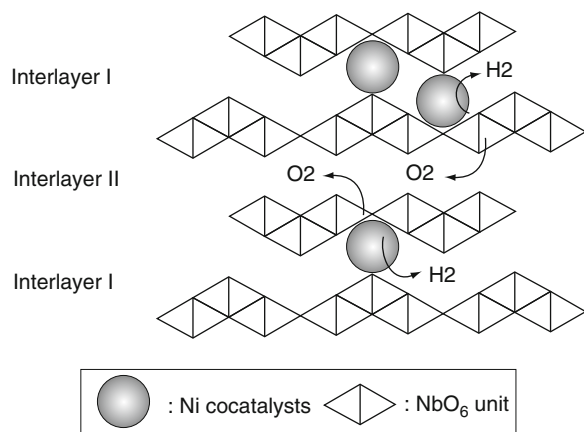
involved are illustrated in Fig. 8. Porphyrins, [Ru(bipy)<sub>3</sub>]<sup>2+</sup>, [Fe(CN)<sub>6</sub>]<sup>4-</sup>, carboxylic (c-RuL<sub>3</sub>) compounds, phosphonic (p-RuL<sub>3</sub>) compounds, diamine and dithiolate complexes of Pt<sup>IV</sup>, eosin, Cu phthalocyanine, and dipyrindyl complexes of Ru, and complex of Zn with cytochrome C are commonly used as dye sensitizers. It is not easy to find a stable dye without any degradation by sensitized photocatalyst after long time irradiation. The dye should also have a strong absorption in visible light and a suitable energy level

of its excited state (more negative than the conduction band of the sensitized semiconductor).

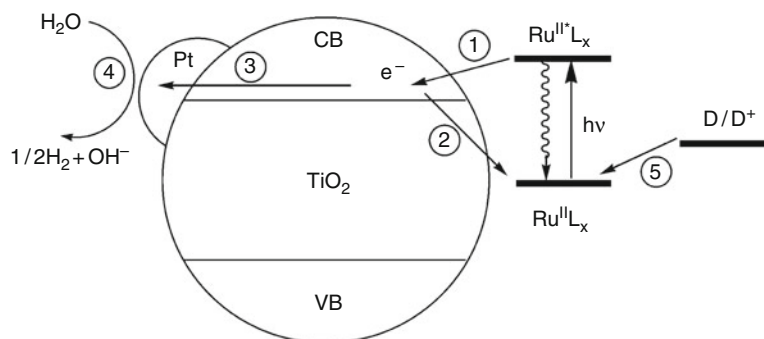


### Cocatalyst Deposited Semiconductor

Cocatalysts, typically noble metal nanoparticles, are normally needed in photocatalytic systems for H<sub>2</sub> production. Apart from noble metals, cocatalyst chemistries also include RuO<sub>2</sub>, Ni and its oxide, the mixed oxides of Rh and Cr, tungsten carbide, MoS<sub>2</sub>, and so on. The promoting mechanism by cocatalysts is illustrated in Fig. 1. When a semiconductor is excited by light irradiation with enough energy, photogenerated electrons are transitioned from the valence band to the conduction band of the semiconductor, raising the Fermi level of the semiconductor. The Fermi level difference between the semiconductor and the deposited noble metal drives the photogenerated electrons to the noble metal, and this increases the Fermi level of the noble metal, which drives the electrons from the noble metal to an electron acceptor in the solution (H<sup>+</sup> for H<sub>2</sub> production). From this process, the cocatalyst improves photocatalytic H<sub>2</sub> production by acting as (1) an electron sink to separate electrons and holes, reducing their recombination, and (2) an electron



**Photo-catalytic Hydrogen Production. Figure 7** Water splitting over K<sub>4</sub>Nb<sub>6</sub>O<sub>17</sub> photocatalyst with layered structure (From [68]. Reproduced by permission of The Royal Society of Chemistry)



**Photo-catalytic Hydrogen Production. Figure 8**

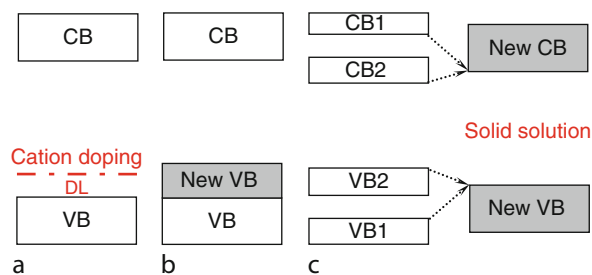
Illustration of visible light-induced H<sub>2</sub> production on a ruthenium complex-sensitized TiO<sub>2</sub> particle in water. The number represents the major electron pathways: 1, electron injection from the excited sensitizer to CB; 2, back electron transfer to the oxidized sensitizer (Ru<sup>III</sup>L<sub>x</sub>); 3, electron migration and trapping in Pt deposits; 4, interfacial electron transfer to H<sub>2</sub>O (or H<sup>+</sup>) on Pt; 5, sensitizer regeneration by an electron donor (D) (Reproduced with permission from [69]. Copyright 2007 Society of Photo Optical Instrumentation Engineers)

transfer to shuttle photogenerated electrons from the semiconductor to an acceptor, reducing the activation energy for the reduction of water. Both the species and size of the noble metal affect the energetics and the electron transfer between the semiconductor and noble metal. From the practical viewpoint, it is economical to exploit and develop cheap noble metal species for deposition modification.

### Doped Semiconductor

Doping has been widely used to extend the optical response of wide band gap semiconductors to visible range. Transition metal doped  $\text{TiO}_2$  was systemically investigated [70]. The photoreactivity of doped  $\text{TiO}_2$  appears to be a complex function of the dopant concentration, the energy level of dopants within the  $\text{TiO}_2$  lattice, their d electronic configuration, the distribution of dopants, the electron donor concentration, and the light intensity. It is important to tune the species, content, depth, and distribution of dopants inside the structure of host photocatalysts. Sometimes, cation-doping-induced visible-light absorption cannot become a substantial condition for photocatalytic activity in the visible range, since the absorption of doped semiconductors results from several absorption transitions of different origins. The doping of wide band gap semiconductors with transition metal ions creates discrete new energy levels within the forbidden band. Visible-light absorption and photoactivity are induced by the interband, as shown in Fig. 9a. This induced visible-light photoactivity is normally low, due to the limited amount of dopants that can be incorporated, as indicated by a small shoulder in the visible-light region, instead of a total red shift of the absorption edge. Increasing the dopant concentration in the semiconductor matrix can, to some extent, improve visible-light absorption, but excessive doping can easily disturb the original structure, or form separate impurity phases. Sometimes, doping can extend the lifetime of photogenerated charge carriers by transient and shallow charge carrier trapping. In most cases, dopants can also act as recombination centers for photogenerated electrons and holes and decrease photocatalytic activity.

Anion doping to improve photocatalytic  $\text{H}_2$  production under visible-light irradiation is a quite new



### Photo-catalytic Hydrogen Production. Figure 9

Existing band engineering approaches: (a) cation doping, which creates a discrete impurity energy level (*DL*) within the forbidden band gap; (b) valence band modification, which forms a new valence band with higher top; and (c) solid solution formation, producing a new couple of valence and conduction bands, whose band gap is between those of the component semiconductors (Reprinted from [1], Copyright (2009). With permission from Elsevier)

field, compared to transition metal cation doping. Although there has been some work on anion doping, the state of anion dopants, and the origin of visible-light absorption and photoactivity are still in the debate. Among all the anion dopants (N, F, C, S, and P), N-doped  $\text{TiO}_2$  has been mostly investigated as a representative, since N has the similar atomic radius and chemical state with O, which makes it easy to substitute O in metal oxide lattice. N-doped  $\text{TiO}_2$  and its visible-light response were first reported by Asahi et al. in 2001 [71], and this work has been followed by various theoretical and experimental studies. Asahi et al. suggested that the N 2p level (above the O 2p level in  $\text{TiO}_2$ ) could mix with the valence band of  $\text{TiO}_2$  composed of O 2p orbits, which forms a new valence band and narrows the band gap (Fig. 9b). The conduction band of N-doped  $\text{TiO}_2$  remains unchanged and higher than the  $\text{H}_2$  reduction potential, while the valence band is shifted up, but still enough for water oxidation. However, this mechanism was challenged by arguing that a low level of doping ( $\leq 2$  at.%) can only form midgap state above the valence band of  $\text{TiO}_2$ , like cation doping, and cannot shift up the valence band unless a high level of doping ( $\geq 20\%$ ) is carried out. In practice, high-level doping may form oxynitrides (or oxysulfides) or even nitrides (or sulfides). It was also argued that the visible-light response is due to the

advent of color centers (e.g., F, F<sup>+</sup>, F<sup>2+</sup>, and Ti<sup>3+</sup>), and the formation of midgap energy level and oxygen vacancies induced by doping [72, 73].

### Solid Solution Semiconductor (Semiconductor Alloy)

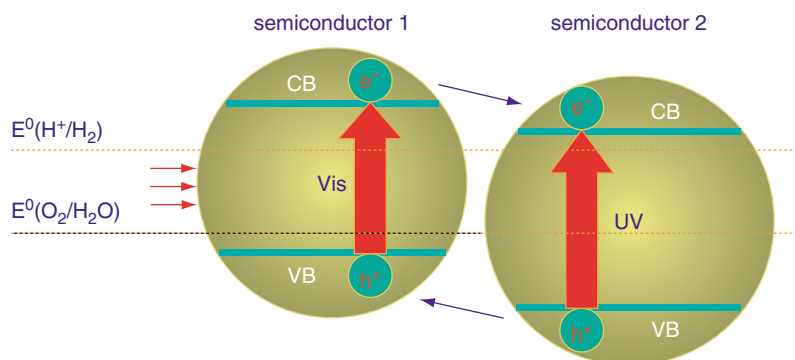
Another approach to adjust the band gap and band positions is to form a solid solution semiconductor (or a semiconductor alloy) between wide and narrow band gap semiconductors (Fig. 9c). If two (or more) semiconductors have similar crystal lattice structures, solid solution is likely formed by them. The band gap of solid solution can be tuned, to some extent, by varying the composition of the solid solution. The reported solid solutions include GaN-ZnO, ZnS-CdS, ZnS-AgInS<sub>2</sub>, ZnS-AgInS<sub>2</sub>-CuInS<sub>2</sub>, CdS-CdSe, ZnO-ZnGeN<sub>2</sub>, GaP-InP, and others. Interestingly, despite the large band gaps of pure GaN and ZnO (>3 eV), GaN-ZnO solid solutions have visible absorption with band gaps of 2.4–2.8 eV, and can decompose water under visible light. Density functional theory (DFT) calculations indicated that the conduction band bottom of GaN-ZnO was mainly composed of 4s and 4p orbitals of Ga, while the valence band top consisted of N 2p orbitals, followed by Zn 3d orbitals. The presence of Zn 3d and N 2p electrons in the upper valence band might provide p-d repulsion for extending the valence band, narrowing the band gap [34].

### Semiconductor Composites

Photogenerated charge carrier recombination can be minimized by coupling two semiconductors, if their band positions are crossed, as illustrated in Fig. 10. Photogenerated holes tend to accumulate in the semiconductor with the less positive valence band (semiconductor 1 in Fig. 10), while photogenerated electrons are collected by the semiconductor with the less negative conduction band (semiconductor 2 in Fig. 10). This efficient charge separation largely enhances the photocatalytic efficiency. It is very useful to choose a narrow band gap semiconductor as one coupling semiconductor in order to utilize visible light. Semiconductors, such as CdS, PbS, Bi<sub>2</sub>S<sub>3</sub>, CdSe, InP, and n-Si, that can absorb visible light serve as inorganic sensitizers in semiconductor/semiconductor composites. Photocatalytic H<sub>2</sub> production by semiconductor composites can be successfully achieved if the following conditions are met: (1) the band levels of two semiconductors should be matched well, leading to wide charge carrier separation, (2) the less negative conduction band of the composite should still be more negative than the water reduction potential, and (3) there should be good contact between the two semiconductors, which ensures fast and efficient charge carrier injection between them.

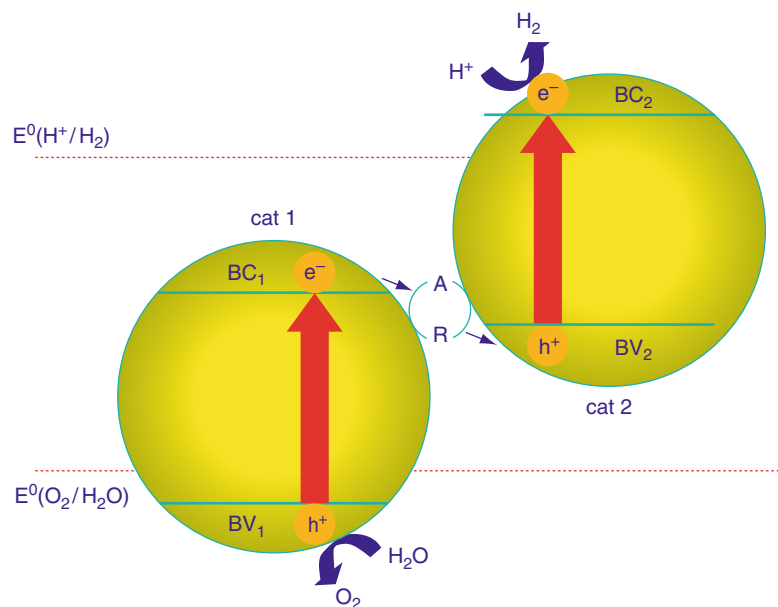
### Multiphotonic System (Z-Scheme) for H<sub>2</sub> Production

By mimicking the natural photosynthesis by green plants, a Z-scheme photocatalytic water splitting



**Photo-catalytic Hydrogen Production. Figure 10**

Band structure of a composite photocatalyst with an enhanced visible-light response, prepared by a mixture of wide and narrow band gap photocatalysts (Reproduced with permission from [74]. Copyright 2009 Wiley-VCH Verlag GmbH & Co. KGaA)



**Photo-catalytic Hydrogen Production. Figure 11**

Diagram of a dual photocatalyst system employing a redox shuttle (A/R) (Reproduced with permission from [74]. Copyright 2009 Wiley-VCH Verlag GmbH & Co. KGaA)

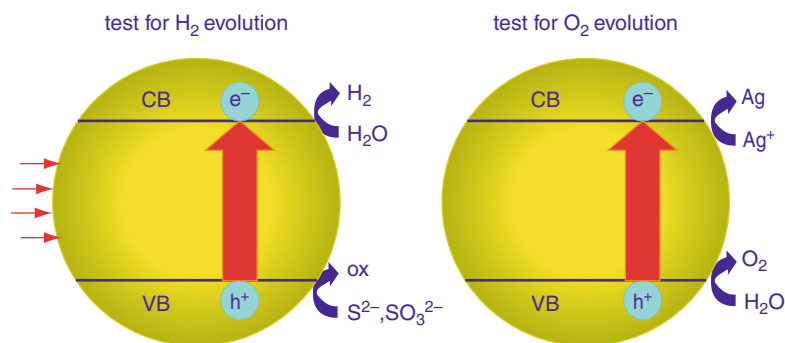
system was developed [36]. This system involved two-photon-excitation under visible light. The Z-scheme system consisted of a H<sub>2</sub>-evolution photocatalyst (Cat 2 in Fig. 11), an O<sub>2</sub>-evolution photocatalyst (Cat 1 in Fig. 11), and a reversible redox mediator (Ox/Red) that acted separately as the electron donor (R in Fig. 11) and acceptor (A in Fig. 11) for the respective half reactions. Protons are reduced to hydrogen molecules by the conduction band electrons of the photocatalyst with the more negative conduction band level, and R scavenges its valence band holes. Water is oxidized to oxygen by the photocatalyst with the more positive valence band level, and A reacts with its conduction band electrons. Photocatalysts, which only work in half reactions in water splitting, combine with each other to run the whole water splitting in this system. These photocatalysts are partly free of the strict energy limitations for the band structure of a single water splitting photocatalyst. They can have narrow band gaps with only one band (conduction or valence band) level enough for water reduction or oxidation. SrTiO<sub>3</sub>, TaON, CaTaO<sub>2</sub>N, and BaTaO<sub>2</sub>N can work as H<sub>2</sub> evolution photocatalysts, while WO<sub>3</sub>, BiVO<sub>4</sub>, and Bi<sub>2</sub>MoO<sub>6</sub> can act as O<sub>2</sub> evolution photocatalysts. The IO<sub>3</sub><sup>-</sup>/I<sup>-</sup>,

Fe<sup>3+</sup>/Fe<sup>2+</sup>, and Ce<sup>4+</sup>/Ce<sup>3+</sup> redox couples normally act as reversible electron shuttles. The key factors for designing a good Z-scheme system are to find a pair of photocatalysts for separate H<sub>2</sub> and O<sub>2</sub> production with high efficiency, and an efficient reversible electron mediator, the redox potential of which can meet the energy requirements of being electron donor and acceptor in the respective half reactions.

#### Dissolved Additives for H<sub>2</sub> Production Improvement

Photocatalytic H<sub>2</sub> or O<sub>2</sub> production is often carried out in the presence of electron donors including low aliphatic alcohols (methanol, ethanol, isopropanol), sulfides (H<sub>2</sub>S, Na<sub>2</sub>S), sulfites (Na<sub>2</sub>SO<sub>3</sub>), hydrazine, aliphatic amines (triethylamine, triethanolamine), carboxylic acids (formic acid, EDTA), carbohydrates and other organic compounds, or electron acceptors including persulfate, Ag<sup>+</sup>, and Fe<sup>3+</sup>. The mechanism of photocatalytic reaction using these sacrificial reagents is illustrated in Fig. 12.

Taking photocatalytic H<sub>2</sub> production for example, after the excitation of the photocatalyst by light with enough energy, the photogenerated electrons in the conduction band reduce absorbed water molecules



**Photo-catalytic Hydrogen Production. Figure 12**

Half reactions of water splitting for H<sub>2</sub> and O<sub>2</sub> evolution reactions in the presence of sacrificial reagents (Reproduced with permission from [74]. Copyright 2009 Wiley-VCH Verlag GmbH & Co. KGaA)

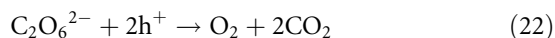
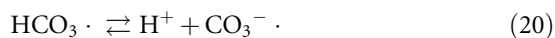
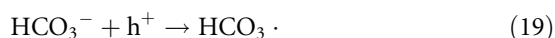
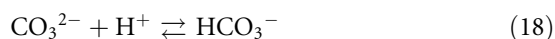
to H<sub>2</sub>, while a hole in the valence band is filled with an electron from a sacrificial reagent (S<sup>2-</sup>, SO<sub>3</sub><sup>2-</sup> in Fig. 12). The thermodynamic requirement for the occurrence of such a process is the potential of the conduction band more negative than the water reduction potential, and the potential of the valence band more positive than the oxidation potential of the electron donor, rather than water. This requirement sometimes does not need strong oxidation ability of photocatalyst, depending on the reducing reagent used.

Although these half reactions are not overall water splitting reaction, they are very useful to test a given semiconductor's kinetic and thermodynamic performance for H<sub>2</sub> or O<sub>2</sub> evolution. Scavengers, e.g., S<sup>2-</sup>/SO<sub>3</sub><sup>2-</sup>, are commonly used when hydrogen production is performed by metal sulfides. When metal sulfides are used for photocatalytic hydrogen production, photo-corrosion occurs due to the S<sup>2-</sup> in metal sulfides being oxidized by photogenerated holes in the valence band. As a sacrificial reagent, S<sup>2-</sup> can easily react with two holes to form S. Then, the added SO<sub>3</sub><sup>2-</sup> can dissolve S into S<sub>2</sub>O<sub>3</sub><sup>2-</sup>, preventing any detrimental deposition of S on CdS. Therefore, photo-corrosion of CdS is avoided. Interestingly, a process of simultaneous hydrogen production and H<sub>2</sub>S removal was performed by a composite photocatalyst made of bulk CdS decorated with TiO<sub>2</sub> nanoparticles, i.e., CdS (bulk)/TiO<sub>2</sub>, under visible light. Hydrogen originated from both H<sub>2</sub>S and H<sub>2</sub>O when H<sub>2</sub>S was dissolved in alkaline water [75].

Since electron donors are sacrificially consumed during photocatalytic H<sub>2</sub> production, continuous

addition of electron donors is normally necessary for sustaining H<sub>2</sub> production. The effect from different hydrocarbon electron donors was qualitatively investigated. It was found that the decomposition of these hydrocarbons could also contribute to a higher H<sub>2</sub> yield, since H<sub>2</sub> is one of their decomposition products [76]. The integration of clean H<sub>2</sub> fuel production and pollutant decomposition is promising and significant in practical application, since some pollutants (oxalic acid, formic acid, and formaldehyde) can act as electron donors [77].

Addition of carbonate salts was found to enhance H<sub>2</sub> and O<sub>2</sub> production stoichiometrically [78]. Several carbonate species were formed through the following reactions (Eqs. 18–21). Photogenerated holes were consumed by reacting with HCO<sub>3</sub><sup>-</sup> to form carbonate radicals (HCO<sub>3</sub><sup>·</sup>, CO<sub>3</sub><sup>·-</sup>) (Eqs. 19, 20), which is beneficial for photogenerated charge carrier separation. Peroxycarbonates (C<sub>2</sub>O<sub>6</sub><sup>2-</sup>) could be easily decomposed by holes into O<sub>2</sub> and CO<sub>2</sub> (Eq. 22). The evolution of CO<sub>2</sub> promoted the desorption of O<sub>2</sub> from the photocatalyst surface, and thus minimized the back reaction between H<sub>2</sub> and O<sub>2</sub>. Then desorbed CO<sub>2</sub> could be dissolved and converted into HCO<sub>3</sub><sup>-</sup> again.



Addition of iodide was also found to promote hydrogen production [79]. Iodide anion ( $I^-$ ) in a suspension can be adsorbed preferentially onto cocatalyst surface, forming an iodine layer. The iodine layer can suppress backward reaction between  $H_2$  and  $O_2$ , which enhanced the production of hydrogen and oxygen significantly. However, adding too much carbonate salts or iodide anions could make these species excessively enriched on the photocatalyst surface, decreasing light harvesting.

### Future Directions

Worldwide energy consumption and man-made global warming result from humans excessively using limited fossil fuel on the earth, so it is time for society to pursue sustainable, clean energy systems. It is likely that the topic of this entry (photocatalytic  $H_2$  production) will contribute greatly to a sustainable, clean energy system, but with uncertain timing for deployment. Although the efficiency of photocatalytic  $H_2$  production is lower than that of photoelectrochemical (PEC) water splitting, the former technology provides a simple and convenient manipulation platform and valuable information for the development of the latter. There are also other promising technologies for sustainable, clean energy systems, which should also be paid attention to, and can be studied together with photocatalytic and PEC water splitting. Photocatalytic fuel generation from  $CO_2$  is one of them. It converts  $CO_2$  and  $H_2O$  to chemical fuels (e.g., methanol) by using solar energy, which decreases global emission of  $CO_2$  and keeps global neutral carbon cycle. It allows, to some extent, human beings to keep using fossil fuel. This technology has many similarities with photocatalytic  $H_2$  production, e.g., water oxidation is also evolved in its process. The research outcomes from photocatalytic  $H_2$  production may well be applicable to the rising technology of photocatalytic  $CO_2$  fixation.

Although photocatalytic water splitting ( $H_2$  production) has been studied for almost 40 years, a further understanding of this process has dropped behind the reported phenomena. Although this field has seen some achievements, the total solar-to-hydrogen conversion efficiency is still much lower (about 1%) than the expected (about 10%) for practical application. Fundamental research on scientific details is necessary.

For instance, is the charge carrier recombination so fast that too few reductive and oxidative sites can be reached by water? Where are the surface-active sites? What is their nature? Does desorption of products mainly determine the low efficiency? Besides studying the mechanism, the criteria for evaluation of photocatalytic  $H_2$  production should be established, which makes the comparison of results from different labs much easier. This kind of criteria has already been set up in the solar cell field. Material development can strongly promote the breakthroughs in this technology. Some traditional semiconductors (metal oxides, metal sulfides) should be modified, rather than discarded, since they have their potential advantages in practical application. Of course, the study of their modification is always important.

Simultaneously, there is an increasing opportunity to exploit novel photocatalysts for  $H_2$  production with the accumulation of theoretical and experimental knowledge. The focus should still be on stable, low-cost photocatalysts that can be manufactured on a large scale. Since using sunlight for  $H_2$  production is a trend or target, the sensitization of photocatalysts and multiphotonic systems are expected topics in this field. Nanoscience and nanotechnology will continue to play an important role in many aspects, including the preparation and characterization of photocatalysts, the modification of optical and electronic structures of photocatalysts, and the design for light harvesting and management. Nanoreactors and well-controlled model systems are expected to be very useful for better understanding of the basic physics and chemistry involved in photocatalytic  $H_2$  production. Further down the road, there will be a significant engineering challenge to integrate individual components into a fully functional device, the scalability and sustainability of which should also be achieved.

### Bibliography

#### Primary Literature

1. Zhu J, Zäch M (2009) Nanostructured materials for photocatalytic hydrogen production. *Curr Opin Colloid Interface Sci* 14:260–269
2. Kamat PV (2007) Meeting the clean energy demand: nanostructure architectures for solar energy conversion. *J Phys Chem C* 111:2834–2860



- Fujishima A, Honda K (1972) Electrochemical photolysis of water at a semiconductor electrode. *Nature* 238:37–38
- Domen K, Naito S, Onishi T, Tamaru T, Soma M (1982) Study of the photocatalytic decomposition of water vapor over a nickel (II) oxide-strontium titanate ( $\text{SrTiO}_3$ ) catalyst. *J Phys Chem* 86:3657–3661
- Domen K, Kudo A, Onishi T, Kosugi N, Kuroda H (1986) Photocatalytic decomposition of water into hydrogen and oxygen over nickel (II) oxide-strontium titanate ( $\text{SrTiO}_3$ ) powder. 1. Structure of the catalysts. *J Phys Chem* 90:292–295
- Matsumura M, Saho Y, Tsubomura H (1983) Photocatalytic hydrogen production from solutions of sulfite using platinized cadmium sulfide powder. *J Phys Chem* 87:3807–3808
- Kakuta N, Park KH, Finlayson MF, Ueno A, Bard AJ, Campion A, Fox MA, Webber SE, White JM (1985) Photoassisted hydrogen production using visible light and coprecipitated  $\text{ZnS-CdS}$  without a noble metal. *J Phys Chem* 89:732–734
- Reber JF, Rusek M (1986) Photochemical hydrogen production with platinized suspensions of cadmium sulfide and cadmium zinc sulfide modified by silver sulfide. *J Phys Chem* 90:824–834
- Reber JF, Meier K (1984) Photochemical production of hydrogen with zinc sulfide suspensions. *J Phys Chem* 88:5903–5913
- Kudo A, Tanaka A, Domen K, Maruya K, Aika K, Onishi T (1988) Photocatalytic decomposition of water over  $\text{NiO-K}_4\text{Nb}_6\text{O}_{17}$  catalyst. *J Catal* 111:67–76
- Kudo A, Sayama K, Tanaka A, Asakura K, Domen K, Maruya K, Onishi T (1989) Nickel-loaded  $\text{K}_4\text{Nb}_6\text{O}_{17}$  photocatalyst in the decomposition of  $\text{H}_2\text{O}$  into  $\text{H}_2$  and  $\text{O}_2$ : structure and reaction mechanism. *J Catal* 120:337–352
- Domen K, Kudo A, Tanaka A, Onishi T (1990) Overall photodecomposition of water on a layered niobate catalyst. *Catal Today* 8:77–84
- Yoshimura J, Kudo A, Tanaka A, Domen K, Maruya K, Onishi T (1988)  $\text{H}_2$  evolution caused by electron transfer between different semiconductors under visible light irradiation. *Chem Phys Lett* 147:401–404
- Shibata M, Kudo A, Tanaka A, Domen K, Maruya K, Onishi T (1985) Photocatalytic activities of layered titanium compounds and their derivatives for  $\text{H}_2$  evolution from aqueous methanol solution. *Chem Lett* 16:1017–1018
- Kato H, Kudo A (1998) New tantalate photocatalysts for water decomposition into  $\text{H}_2$  and  $\text{O}_2$ . *Chem Phys Lett* 295:487–492
- Mitsui C, Nishiguchi H, Fukamachi K, Ishihara T, Takita Y (1999) Photocatalytic decomposition of pure water over  $\text{NiO}$  supported on  $\text{KTa(M)O}_3$  ( $\text{M} = \text{Ti}^{4+}, \text{Hf}^{4+}, \text{Zr}^{4+}$ ) perovskite oxide. *Chem Lett* 28:1327–1328
- Ishihara T, Nishiguchi H, Fukamachi K, Takita Y (1999) Effects of acceptor doping to  $\text{KTaO}_3$  on photocatalytic decomposition of pure  $\text{H}_2\text{O}$ . *J Phys Chem B* 103:1–3
- Kato H, Kudo A (1999) Photocatalytic decomposition of pure water into  $\text{H}_2$  and  $\text{O}_2$  over  $\text{SrTa}_2\text{O}_6$  prepared by a flux method. *Chem Lett* 28:1207–1208
- Machida M, Yabunaka J, Kijima T (1999) Efficient photocatalytic decomposition of water with the novel layered tantalate  $\text{RbNdTa}_2\text{O}_7$ . *Chem Commun* 1999:1939–1940
- Kudo A, Hiji S (1999)  $\text{H}_2$  or  $\text{O}_2$  evolution from aqueous solutions on layered oxide photocatalysts consisting of  $\text{Bi}^{3+}$  with  $6s^2$  configuration and  $d^0$  transition metal ions. *Chem Lett* 28:1103–1104
- Kudo A, Kato H (1997) Photocatalytic activities of  $\text{Na}_2\text{W}_4\text{O}_{13}$  with layered structure. *Chem Lett* 26:421–422
- Sato J, Saito S, Nishiyama H, Inoue Y (2001) New photocatalyst group for water decomposition of  $\text{RuO}_2$ -loaded p-block metal (In, Sn, and Sb) oxides with  $d^{10}$  configuration. *J Phys Chem B* 105:6061–6063
- Ikarashi K, Sato J, Kobayashi H, Saito S, Nishiyama H, Inoue Y (2002) Photocatalysis for water decomposition by  $\text{RuO}_2$ -dispersed  $\text{ZnGa}_2\text{O}_4$  with  $d^{10}$  configuration. *J Phys Chem B* 106:9048–9053
- Sato J, Saito S, Nishiyama H, Inoue Y (2003) Photocatalytic activity for water decomposition of indates with octahedrally coordinated  $d^{10}$  configuration. I. Influences of preparation conditions on activity. *J Phys Chem B* 107:7965–7969
- Sato J, Kobayashi H, Inoue Y (2003) Photocatalytic activity for water decomposition of indates with octahedrally coordinated  $d^{10}$  configuration. II. Roles of geometric and electronic structures. *J Phys Chem B* 107:7970–7975
- Sato J, Kobayashi H, Ikarashi K, Saito S, Nishiyama H, Inoue Y (2004) Photocatalytic activity for water decomposition of  $\text{RuO}_2$ -dispersed  $\text{Zn}_2\text{GeO}_4$  with  $d^{10}$  configuration. *J Phys Chem B* 108:4369–4375
- Sato J, Saito N, Yamada Y, Maeda K, Takata T, Kondo JN, Hara M, Kobayashi H, Domen K, Inoue Y (2005)  $\text{RuO}_2$ -loaded  $\beta\text{-Ge}_3\text{N}_4$  as a non-oxide photocatalyst for overall water splitting. *J Am Chem Soc* 127:4150–4151
- Hara M, Takata T, Kondo JN, Domen K (2004) Photocatalytic reduction of water by  $\text{TaON}$  under visible light irradiation. *Catal Today* 90:313–317
- Yamasita D, Takata T, Hara M, Kondo JN, Domen K (2004) Recent progress of visible-light-driven heterogeneous photocatalysts for overall water splitting. *Solid State Ionics* 172:591–595
- Kasahara A, Nukumizu K, Takata T, Kondo JN, Hara M, Kobayashi H, Domen K (2003)  $\text{LaTiO}_2\text{N}$  as a visible-light ( $\leq 600$  nm)-driven photocatalyst (2). *J Phys Chem B* 107:791–797
- Liu M, You W, Lei Z, Zhou G, Yang J, Wu G, Ma G, Luan G, Takata T, Hara M, Domen K, Li C (2004) Water reduction and oxidation on  $\text{Pt-Ru/Y}_2\text{Ta}_2\text{O}_5\text{N}_2$  catalyst under visible light irradiation. *Chem Commun* 2004:2192–2193
- Ishikawa A, Takata T, Kondo JN, Hara M, Kobayashi H, Domen K (2002) Oxysulfide  $\text{Sm}_2\text{Ti}_2\text{S}_2\text{O}_5$  as a stable photocatalyst for water oxidation and reduction under visible light irradiation ( $\lambda \leq 650$  nm). *J Am Chem Soc* 124:13547–13553
- Ishikawa A, Takata T, Matsumura T, Kondo JN, Hara M, Kobayashi H, Domen K (2004) Oxysulfides  $\text{Ln}_2\text{Ti}_2\text{S}_2\text{O}_5$  as stable photocatalysts for water oxidation and reduction under visible-light irradiation. *J Phys Chem B* 108:2637–2642
- Maeda K, Teramura K, Lu DL, Takata T, Saito N, Inoue Y, Domen K (2006) Photocatalyst releasing hydrogen from water-Enhancing catalytic performance holds promise for

- hydrogen production by water splitting in sunlight. *Nature* 440:295
35. Lee Y, Terashima H, Shimodaira Y, Teramura K, Hara M, Kobayashi H, Domen K, Yashima M (2007) Zinc Germanium Oxynitride as a photocatalyst for overall water splitting under visible light. *J Phys Chem C* 111:1042–1048
  36. Sayama K, Mukasa K, Abe R, Abe Y, Arakawa H (2002) A new photocatalytic water splitting system under visible light irradiation mimicking a Z-scheme mechanism in photosynthesis. *J Photochem Photobiol A Chem* 148:71–77
  37. Tada H, Mitsui T, Kiyonaga T, Akita T, Tanaka K (2006) All-solid-state Z-scheme in CdS–Au–TiO<sub>2</sub> three-component nanojunction system. *Nat Mater* 5:782–786
  38. Parmon V, Emeline AV, Serpone N (2002) Glossary of terms in photocatalysis and radiocatalysis. *Int J Photoenergy* 4:91–131
  39. Kim HG, Hwang DW, Kim J, Kim YG, Lee JS (1999) Highly donor-doped (110) layered perovskite materials as novel photocatalysts for overall water splitting. *Chem Commun* 1999:1077–1078
  40. Kato H, Kudo A (2001) Water splitting into H<sub>2</sub> and O<sub>2</sub> on alkali tantalate photocatalysts ATaO<sub>3</sub> (A = Li, Na, and K). *J Phys Chem B* 105:4285–4292
  41. Shangquan WF (2007) Hydrogen evolution from water splitting on nanocomposite photocatalysts. *Sci Tech Adv Mater* 8:76–81
  42. Hitoki G, Ishikawa A, Takata T, Kondo JN, Hara M, Domen K (2002) Ta<sub>3</sub>N<sub>5</sub> as a novel visible light-driven photocatalyst ( $\lambda < 600$  nm). *Chem Lett* 31:736–737
  43. Kida T, Minami Y, Guan G, Nagano M, Akiyama M, Yoshida A (2006) Photocatalytic activity of gallium nitride for producing hydrogen from water under light irradiation. *J Mater Sci* 41:3527–3534
  44. Heller A (1984) Hydrogen-evolving solar cells. *Science* 223:1141–1148
  45. Khaselev O, Turner JA (1998) A monolithic photoelectrochemical device for hydrogen production via water splitting. *Science* 280:425–427
  46. Yi H, Peng T, Ke D, Ke D, Zan L, Yan C (2008) Photocatalytic H<sub>2</sub> production from methanol aqueous solution over titania nanoparticles with mesostructures. *Int J Hydrogen Energy* 33:672–678
  47. Lakshminarasimhan N, Bae E, Choi W (2007) Enhanced photocatalytic production of H<sub>2</sub> on mesoporous TiO<sub>2</sub> prepared by template-free method: role of interparticle charge transfer. *J Phys Chem C* 111:15244–15250
  48. Korzhak AV, Ermokhina NI, Stroyuk AL, Bukhtiyarov VK, Raevskaya AE, Litvin VI, Kuchmij SY, Ilyin VG, Manorik PA (2008) Photocatalytic hydrogen evolution over mesoporous TiO<sub>2</sub>/metal nanocomposites. *J Photochem Photobiol A Chem* 198:126–134
  49. Zhang YJ, Zhang L (2008) Synthesis of composite material CdS/Al-HMS and hydrogen production by photocatalytic pollutant degradation under visible light irradiation. *J Inorg Mater* 23:66–70
  50. Ryu SY, Balcerski W, Lee TK, Hoffmann MR (2007) Photocatalytic production of hydrogen from water with visible light using hybrid catalysts of CdS attached to microporous and mesoporous silicas. *J Phys Chem C* 111:18195–18203
  51. Lunawat PS, Senapati S, Kumar R, Gupta NM (2007) Visible light-induced splitting of water using CdS nanocrystallites immobilized over water-repellant polymeric surface. *Int J Hydrogen Energy* 32:2784–2790
  52. Guan GQ, Kida T, Kusakabe K, Kimura K, Fang XM, Ma TL, Abe E, Yoshida A (2004) Photocatalytic H<sub>2</sub> evolution under visible light irradiation on CdS/ETS-4 composite. *Chem Phys Lett* 385:319–322
  53. Beerermann N, Vayssieres L, Lindquist S-E, Hagfeldt A (2000) Photoelectrochemical studies of oriented nanorod thin films of hematite. *J Electrochem Soc* 147:2456–2461
  54. van de Krol R, Liang Y, Schoonman J (2008) Solar hydrogen production with nanostructured metal oxides. *J Mater Chem* 18:2311–2320
  55. Lin CH, Lee CH, Chao JH, Kuo CY, Cheng YC, Huang WN, Chang HW, Huang YM, Shih MK (2004) Photocatalytic generation of H<sub>2</sub> gas from neat ethanol over Pt/TiO<sub>2</sub> nanotube catalysts. *Catal Lett* 98:61–66
  56. Nam W, Han GY (2007) Preparation and characterization of anodized Pt-TiO<sub>2</sub> nanotube arrays for water splitting. *J Chem Eng Jpn* 40:266–269
  57. Khan MA, Akhtar MS, Woo SI, Yang OB (2008) Enhanced photoresponse under visible light in Pt ionized TiO<sub>2</sub> nanotube for the photocatalytic splitting of water. *Catal Commun* 10:1–5
  58. Kuo HL, Kuo CY, Liu CH, Chao JH, Lin CH (2007) A highly active bi-crystalline photocatalyst consisting of TiO<sub>2</sub> (B) nanotube and anatase particle for producing H<sub>2</sub> gas from neat ethanol. *Catal Lett* 113:7–12
  59. Thimsen E, Rastgar N, Biswas P (2008) Nanostructured TiO<sub>2</sub> films with controlled morphology synthesized in a single step process: performance of dye-sensitized solar cells and photo watersplitting. *J Phys Chem C* 112:4134–4140
  60. Jitputti J, Suzuki Y, Yoshikawa S (2008) Synthesis of TiO<sub>2</sub> nanowires and their photocatalytic activity for hydrogen evolution. *Catal Commun* 9:1265–1271
  61. Lin CH, Chao JH, Liu CH, Chang JC, Wang FC (2008) Effect of calcination temperature on the structure of a Pt/TiO<sub>2</sub> (B) nanofiber and its photocatalytic activity in generating H<sub>2</sub>. *Langmuir* 24:9907–9915
  62. Janet CM, Viswanath RP (2006) Large scale synthesis of CdS nanorods and its utilization in photo-catalytic H<sub>2</sub> production. *Nanotechnology* 17:5271–5277
  63. Jang JS, Joshi UA, Lee JS (2007) Solvothermal synthesis of CdS nanowires for photocatalytic hydrogen and electricity production. *J Phys Chem C* 111:13280–13287
  64. Yu JG, Qi LF, Jaroniec M (2010) Hydrogen production by photocatalytic water splitting over Pt/TiO<sub>2</sub> nanosheets with exposed (001) facets. *J Phys Chem C* 114:13118–13125

65. Amano F, Prieto-Mahaney OO, Terada Y, Yasumoto T, Shibayama T, Ohtani B (2009) Decahedral single-crystalline particles of anatase titanium(IV) oxide with high photocatalytic activity. *Chem Mater* 21:2601–2603
66. Liu G, Yang HG, Wang XW, Cheng L, Lu H, Wang L, Lu GQ, Cheng HM (2009) Enhanced photoactivity of oxygen-deficient anatase TiO<sub>2</sub> sheets with dominant 001 facets. *J Phys Chem C* 113:21784–21788
67. Youngblood WJ, Lee SHA, Maeda K, Mallouk TE (2009) Visible light water splitting using dye-sensitized oxide semiconductors. *Acc Chem Res* 42:1966–1973
68. Kudo A, Miseki Y (2009) Heterogeneous photocatalyst materials for water splitting. *Chem Soc Rev* 38:253–278
69. Choi W (2007) Photocatalytic hydrogen production using surface modified titania nanoparticles. In: Guo J (ed) *Solar hydrogen and nanotechnology II*. Proc SPIE 6650:66500L
70. Choi W, Termin A, Hoffmann MR (1994) The role of metal-ion dopants in quantum-sized TiO<sub>2</sub>: correlation between photoreactivity and charge-carrier recombination dynamics. *J Phys Chem* 98:13669–13679
71. Asahi R, Morikawa T, Ohwaki T, Aoki K, Taga Y (2001) Visible-light photocatalysis in nitrogen-doped titanium oxides. *Science* 293:269–271
72. Serpone N (2006) Is the band gap of pristine TiO<sub>2</sub> narrowed by anion- and cation-doping of titanium dioxide in second-generation photocatalysts? *J Phys Chem B* 110:24287–24293
73. Fujishima A, Zhang X, Tryk DA (2008) TiO<sub>2</sub> photocatalysis and related surface phenomena. *Surf Sci Rep* 63:515–582
74. Yerga RMN, Galván MCÁ, del Valle F, de la Mano JAV, Fierro JLG (2009) Water splitting on semiconductor catalysts under visible-light irradiation. *ChemSusChem* 2:471–485
75. Jang JS, Kim HG, Borse PH, Lee JS (2007) Simultaneous hydrogen production and decomposition of H<sub>2</sub>S dissolved in alkaline water over CdS-TiO<sub>2</sub> composite photocatalysts under visible light irradiation. *Int J Hydrogen Energy* 32:4786–4791
76. Nada AA, Barakat MH, Hamed HA, Mohamed NR, Veziroglu TN (2005) Studies on the photocatalytic hydrogen production using suspended modified TiO<sub>2</sub> photocatalysts. *Int J Hydrogen Energy* 30:687–691
77. Li YX, Lu GX, Li SB (2003) Photocatalytic production of hydrogen in single component and mixture systems of electron donors and monitoring adsorption of donors by in situ infrared spectroscopy. *Chemosphere* 52:843–850
78. Sayama K, Arakawa H (1992) Significant effect of carbonate addition on stoichiometric photodecomposition of liquid water into hydrogen and oxygen from platinum-titanium (IV) oxide suspension. *J Chem Soc Chem Commun* 2:150–152
79. Abe R, Sayama K, Arakawa H (2003) Significant effect of iodide addition on water splitting into H<sub>2</sub> and O<sub>2</sub> over Pt-loaded TiO<sub>2</sub> photocatalyst: suppression of backward reaction. *Chem Phys Lett* 371:360–364

## Books and Reviews

- Ashokkumar M (1998) An overview on semiconductor particle systems for photoproduction of hydrogen. *Int J Hydrogen Energy* 23:427–438
- Best JP, Dunstan DE (2009) Nanotechnology for photolytic hydrogen production: colloidal anodic oxidation. *Int J Hydrogen Energy* 34:7562–7578
- Fujishima A, Zhang X, Tryk DA (2007) Heterogeneous photocatalysis: from water photolysis to applications in environmental cleanup. *Int J Hydrogen Energy* 32:2664–2672
- Getoff N (1990) Photoelectrochemical and photocatalytic methods of hydrogen production: a short review. *Int J Hydrogen Energy* 15:407–417
- Kaneko M, Okura I (2002) *Photocatalysis: science and technology*. Kodansha/Springer, Tokyo/Berlin
- Kitano M, Hara M (2009) Heterogeneous photocatalytic cleavage of water. *J Mater Chem* 20:627–641
- Kudo A (2007) Photocatalysis and solar hydrogen production. *Pure Appl Chem* 79:1917–1927
- Kudo A (2007) Recent progress in the development of visible light-driven powdered photocatalysts for water splitting. *Int J Hydrogen Energy* 32:2673–2678
- Kudo A, Kato H, Tsuji I (2007) Strategies for the development of visible-light-driven photocatalysts for water splitting. *Chem Lett* 33:1534–1539
- Maeda K, Domen K (2007) New non-oxide photocatalysts designed for overall water splitting under visible light. *J Phys Chem C* 111:7851–7861
- Moon SC, Matsumura Y, Kitano M, Matsuoka M, Anpo M (2003) Hydrogen production using semiconducting oxide photocatalysts. *Res Chem Intermed* 29:233–256
- Navarro RM, Sánchez-Sánchez MC, Alvarez-Galvan MC, del Valle F, Fierro JLG (2009) Hydrogen production from renewable sources: biomass and photocatalytic opportunities. *Energy Environ Sci* 2:35–54
- Osterloh FE (2008) Inorganic materials as catalysts for photochemical splitting of water. *Chem Mater* 20:35–54
- Paleocrassas S (1974) Photocatalytic hydrogen production: a solar energy conversion alternative? *Sol Energy* 16:45–51
- Rajeshwar K, McConnell R, Licht S (2008) *Solar hydrogen generation*. Springer, New York
- Stroyuk AL, Kryukov AI, Kuchmii SY, Pokhodenko VD (2009) Semiconductor photocatalytic systems for the production of hydrogen by the action of visible light. *Theor Exp Chem* 45:209–233
- Zhang H, Chen G, Bahnemann DW (2009) Photoelectrocatalytic materials for environmental applications. *J Mater Chem* 19:5089–5121
- Zäch M, Häggglund C, Chakarov D, Kasemo B (2006) Nanoscience and nanotechnology for advanced energy systems. *Curr Opin Solid State Mater* 10:132–143

## Photosynthetically Active Radiation: Measurement and Modeling

MATTI MÖTTUS<sup>1</sup>, MADIS SULEV<sup>2</sup>, FRÉDÉRIC BARET<sup>3</sup>,  
RAOUL LOPEZ-LOZANO<sup>3</sup>, ANU REINART<sup>2</sup>

<sup>1</sup>Department of Geosciences and Geography,  
University of Helsinki, Helsinki, Finland

<sup>2</sup>Tartu Observatory, Tõravere, Tartumaa, Estonia

<sup>3</sup>INRA, UMR Environnement Méditerranéen et  
Modélisation des Agro-Hydrosystèmes, EMMAH,  
Avignon, France

### Article Outline

Glossary

Definition of the Subject

Introduction

Techniques for Measuring PAR

PAR in Various Environments

PAR in Vegetation Canopies

Future Directions

Bibliography

### Glossary

**Photosynthetically active radiation (PAR)** The part of electromagnetic radiation that can be used as the source of energy for photosynthesis by green plants, measured as PAR irradiance or PPFd.

**PAR waveband** Spectral region for electromagnetic radiation defined by the wavelength limits of 400–700 nm.

**PAR irradiance** Radiant flux density, or the radiative energy received by unit surface area in unit time, carried by photons in the PAR waveband.

**Photosynthetic photon flux density (PPFD)** The number of photons with wavelengths in the PAR waveband passing through unit surface area in unit time; synonymous to PAR quantum flux.

**Photosynthetic action spectrum** The spectral dependence of photosynthetic productivity per unit absorbed energy, usually plotted in relative units.

**IPAR** Intercepted PAR, or the amount of incident PAR not directly transmitted to the ground by a vegetation canopy.

**APAR** Absorbed PAR, the amount of incident PAR absorbed by a vegetation canopy.

**fIPAR** The fraction of incident PAR not directly transmitted to the ground by a vegetation canopy.

**fAPAR** The fraction of incident PAR absorbed by a vegetation canopy.

**Global PAR** The sum of diffuse and direct PAR: total PAR falling on a horizontal surface.

**Ideal PAR energy sensor** PAR sensor with output proportional to PAR irradiance.

**Ideal PAR quantum sensor** PAR sensor with output proportional to PPFd.

**Spectral error** Broadband radiation measurement errors arising from the deviation of the predicted radiation spectrum from the actual one.

**Radiative transfer theory (RTT)** The mathematical framework for describing the radiation field in an absorbing, scattering, and emitting medium based on radiation beams traveling in straight lines.

### Definition of the Subject

In the broad sense, photosynthetically active radiation (PAR) is the part of electromagnetic radiation that can be used as the source of energy for photosynthesis by green plants. Technically, it is defined as radiation in the spectral range from 400 to 700 nm [1, 2]. It is expressed either in terms of photosynthetic photon flux density (PPFD,  $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ ), since photosynthesis is a quantum process, or in terms of photosynthetic radiant flux density (PAR irradiance,  $\text{W m}^{-2}$ ), more suitable for energy balance studies. A fundamental term in the quantification of light used by plants in the photosynthesis process is the fraction of absorbed photosynthetically active radiation (fAPAR) calculated as the ratio of absorbed to total incident PAR in a vegetation canopy. This variable is widely used in vegetation functioning models at a range of spatial scales from the plant to the globe as an indicator of the amount of energy available for photosynthesis [3].

### Introduction

#### Defining PAR

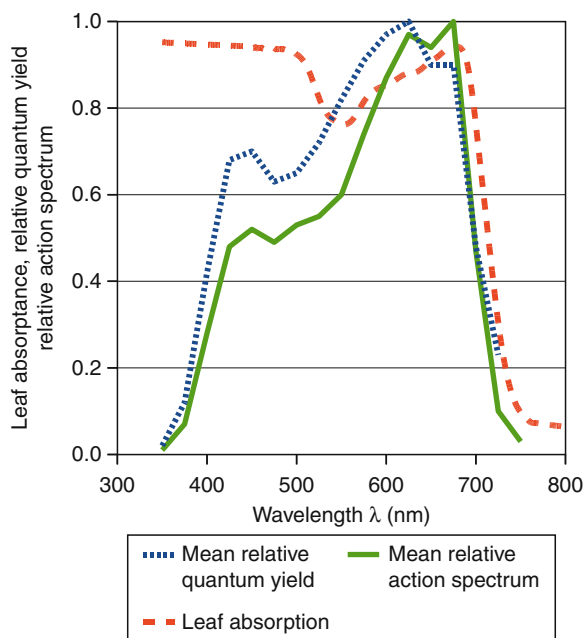
Photosynthetically active radiation (PAR) is commonly defined as electromagnetic radiation in the waveband

between 400 and 700 nm, or 0.400–0.700  $\mu\text{m}$  [1, 2, 4, 5]. The modern definition of PAR arises from the understanding that the measurement system should be based on a single, generalized spectral response curve based on measured data and usable with sufficient accuracy for all practical purposes [6]. This response curve is commonly known as the action spectrum of photosynthetic radiation and is defined as the photosynthetic productivity (measured as  $\text{CO}_2$  uptake or production of  $\text{O}_2$ ) of a leaf plotted against the wavelength  $\lambda$  of the incident spectral irradiance  $I_\lambda$ .

In addition to the action spectrum, the efficiency of photosynthesis is often presented as quantum yield: photosynthetic productivity divided by the amount of absorbed photons. Both quantities are plotted in relative units in Fig. 1: the maximum value of the action spectrum and the relative spectral quantum yield are normalized to unity. The action spectrum and relative spectral quantum yield differ (1) in the units used to measure radiation (amount of photons or amount of radiative energy), and (2) whether the incident or absorbed radiation flux is used. Radiation units for monochromatic radiation can be easily converted: the number of photons with wavelength  $\lambda$  and the corresponding spectral irradiance  $I_\lambda$  are connected via the Planck law (see section “Quantifying PAR”). Absorbed radiative energy (or, equivalently, the number of photons) for each wavelength can be obtained from incident energy by multiplying it by the leaf spectral absorbance.

The shape of photosynthetic action spectrum is almost universal [1, 7]. Small variations are due to between-species differences (e.g., differences in the blue and ultraviolet spectral regions have been noted for arboreous and herbaceous plants [7]), differences in development phase, place of growth, water supply, mineral nutrition, incident irradiance, and other locally varying conditions. This is due to all plants containing the same photochemical apparatus based on the same radiation-absorbing pigments like chlorophyll-A, chlorophyll-B, and carotenoids. These pigments also govern the leaf spectral absorption in the PAR waveband. Only at blue wavelengths, a considerable absorption by non-photosynthetic pigments can be observed [7].

The photosynthetic action spectrum does not decrease to zero at the limits of the PAR waveband



**Photosynthetically Active Radiation: Measurement and Modeling. Figure 1**

The mean spectral absorption of green leaves (average of measurements in Estonia for common local broad-leaf tree species in 2006) together with the action spectrum of PAR and the relative spectral quantum yield of photosynthesis for field-grown plants (Tables IV and VI in [1])

since the change in photosynthetic potential at 400 and 700 nm is fast but not abrupt. Thus, to exactly measure the true photosynthetic potential of incident radiation, one would need to calculate the incident photon flux density weighed by the relative photosynthetic action spectrum for all wavelengths where the action spectrum is not zero, between 360 and 760 nm. The quantity thus obtained is known as the yield photon flux (YPF) [8].

However, to simplify calculations and measurements, the limits of the PAR waveband have been set to 400 and 700 nm by convention, ignoring the relatively small photosynthetic contribution of photons with wavelengths below 400 nm or above 700 nm. Additionally, the true action spectrum and the spectral composition of incident radiation are generally not used, except for most detailed calculations. Instead, an integral value known as PPFD (section “Quantifying PAR”)

is applied (both as a measured value and a theoretical driving force behind photosynthesis in mathematical models) as an adequate descriptor of the photosynthesis-inducing capability of incident radiation under most illumination conditions [2]. The small improvement achievable by using the detailed curve in Fig. 1 does not outweigh the increase in technical and computational complexities.

The radiation incident on a plant canopy arrives as direct and diffuse fluxes. The direct flux is formed by photons having passed through the atmosphere unscattered, whereas the diffuse flux consists of photons scattered by air molecules, aerosol particles, or clouds. As the two fluxes penetrate a vegetation canopy, photons hitting the leaves and other plant elements are intercepted, that is, removed from the incident fluxes. This photon flux hitting plant elements is known as the intercepted PAR flux and denoted with IPAR. Only the intercepted fraction of radiation, or IPAR, constitutes a potential energy source for photosynthesis. However, not all of this potential is realized: a fraction of radiation is always reflected or transmitted by the intercepting element. After being transmitted or reflected, photons may eventually escape the vegetation canopy without any contribution to photosynthesis.

Only photons actually absorbed by the canopy constitute the absorbed PAR (APAR) flux and may be used for photosynthesis. It usually holds that  $APAR < IPAR$  and a constant coefficient,  $APAR = 0.85 IPAR$ , has been proposed for radiation use efficiency calculations [9] based on the work presented in [10]. Both IPAR and APAR are often expressed in relative units as fractional IPAR (fIPAR) and fractional APAR (fAPAR), respectively, by dividing the relevant quantity by the incident PAR flux. These fractional quantities are expressed as numbers between 0 (no interception or no absorption) and 1 (total interception or total absorption). More details on calculating and measuring APAR, fAPAR, and fIPAR are given in section “PAR in Vegetation Canopies”.

### Quantifying PAR

The radiometric quantity for measuring the amount of radiation falling on unit area of a surface (e.g., plant leaf) is irradiance, also known as radiant flux density. The SI (International System of Units) unit for

irradiance is watts per square meter ( $W m^{-2}$ ): thus, electromagnetic radiation is described in terms of the energy it carries. Generally, the term “irradiance” is used to denote the energy carried by photons regardless of their wavelength. When dealing with photons in the PAR waveband, the term “PAR irradiance” should be used to denote the irradiance contributed by photons with wavelengths between 400 and 700 nm:

$$I_{PAR} = \int_{400}^{700} I_{\lambda}(\lambda) d\lambda, \quad (1)$$

where  $I_{\lambda}$  is spectral irradiance.

PAR measurement in plant sciences has aimed at quantitatively describing radiation as the driving force behind photosynthesis. The intensity of photosynthesis is better predicted by the number of absorbed photons than by the radiant energy received by a leaf [2, 6]. This is illustrated by the flatter, more rectangular shape of the quantum yield curve in Fig. 1 compared with that of the action spectrum. For this reason, PAR is often measured as a flux of photons, the quanta of electromagnetic radiation. As we are dealing with the PAR waveband, the particle flux is most commonly termed “photosynthetic photon flux density” or PPF. Mathematically, PPF is defined as the number of photons with wavelengths in the PAR waveband crossing a small surface element in unit time divided by the area of the element.

There is no official SI unit for photon flux or PPF. A unit defined after the famous physicist, Einstein, is used to designate one mole or Avogadro’s number ( $N_A = 6.022 \times 10^{23}$ ) of photons. To describe the PPF under natural illumination conditions, a suitable unit is thus  $\mu E m^{-2} s^{-1}$  (microeinstains per square meter per second). In modern practice, however,  $\mu mol m^{-2} s^{-1}$  (micromoles of photons per square meter per second) is the most extensively used unit for PPF. The increased popularity of micromoles compared with microeinstains is explained by the common requirement of scientific publishers to use SI units whenever possible. The base unit for amount of substance of the international system of units is the mole. Although the (micro) einstein is based on the mole, it is not on the list of SI-derived units. At the same time,  $\mu mol m^{-2} s^{-1}$  is a combination of SI units and thus explicitly compatible with it. Use of  $\mu mol m^{-2} s^{-1}$  for

measuring PAR is also suggested by the International Commission on Illumination [5].

For a monochromatic beam of radiation, the flux of photons is proportional to the flux of energy. The coefficient of proportionality results from Planck law: the energy of a photon is related to its wavelength  $\lambda$  as  $E = hc/\lambda$ , where  $h$  is the Planck constant ( $6.64 \times 10^{-34}$  J s) and  $c$  is the speed of light in vacuum ( $c = 3.00 \times 10^8$  m s<sup>-1</sup>). Thus, we may write the mathematical definition of PPFD as

$$Q_{\text{PAR}} = \int_{400}^{700} \frac{I_{\lambda}(\lambda)}{hcN_A} \lambda d\lambda \quad (2)$$

and define the broadband conversion factor

$$\frac{Q_{\text{PAR}}}{I_{\text{PAR}}} = \frac{1}{hcN_A} \frac{\int_{400}^{700} I_{\lambda}(\lambda) \lambda d\lambda}{\int_{400}^{700} I_{\lambda}(\lambda) d\lambda} \quad (3)$$

For a waveband such as that of PAR containing many wavelengths, the conversion factor  $Q_{\text{PAR}}/I_{\text{PAR}}$  depends on the actual spectral composition of radiation, that is, on irradiance conditions [11]. The technical aspects of the problem are further discussed in section “[Calibration and Spectral Corrections](#)”, and experimental values for PPFD to irradiance conversion are given in section “[PAR Below the Atmosphere](#)”.

There have been several attempts to define PAR in the history of photosynthesis research. Currently, there is very little ambiguity in the term PAR with regard to the wavelength interval. However, other intervals have also been used [12–14], most notably the interval between 380 and 710 nm (e.g., in the Soviet Union, [15]). Thus, historical PAR measurement data may not be compatible with modern data sets despite similar measurement units: careful evaluation and recalibration is required when dealing with long time series.

PAR waveband coincides almost exactly with the visible part of the solar spectrum. The similarity of the wavelength ranges of PAR and visible light may be useful in solving scientific problems. For example, the directional distribution of diffuse sky brightness has been parametrized for different sky conditions [16] and using the high correlation between PAR and visible light, these distributions may be useful in modeling directional distribution of incident PAR. Similarly, an expression such as “availability of light” is reasonable in everyday use. In scientific literature, however, the less

ambiguous term “radiation” should always be preferred to “light.” Ambiguity may emerge as the science of visible light, photometry, has long traditions in quantitative measurements: a standardized luminosity function is used to describe the brightness of radiation as perceived by the human eye. The luminosity function is analogous to the action spectrum of PAR in defining the response of a biological system (the average human eye) to incident radiation. Photometric units have a strong user base with a wide field of applications. The modern unit for measuring visible light incident on a surface (illuminance), lux, was widely used in photosynthesis research half a century ago. However, despite the similarity of the luminosity function and the action spectrum, human vision is not related to the photobiology of photosynthesis, and the use of photometric units and terminology in treatment of PAR is strongly discouraged.

### Fundamentals of Radiation Transfer Theory

The physical laws and concepts used in describing the complex interactions of electromagnetic waves with matter can be readily applied to describe the processes related to PAR. However, trying to follow this path would ultimately lead to tracking every wave or particle using quantum electrodynamics. While accurate laws are used, for example, to describe the scattering of radiation (including PAR) by molecules, aerosols, and cloud particles in the atmosphere, it is impractical to apply the fundamental theory to plants, vegetation canopies, or the whole planet. The common formulation used for accurate computations of PAR, called the radiative transfer theory (RTT), is a simplification based on ray optics: radiation is described in terms of photon bundles traveling in straight lines with infinite velocity.

In principle, RTT is a mathematical formulation of the law of conservation of radiative energy. Using given sources of radiation and the absorptive, scattering, and emissive properties of the medium, it predicts the detailed angular and spatial distribution of radiation. RTT describes radiation in terms of the energy it carries. However, since it is defined for monochromatic radiation, the particle and energy flows are proportional. RTT is a special case of the more general transfer theory dealing with particles (e.g., neutrons or

electrons) in a scattering, absorbing, and generating medium.

Radiative transfer theory is exactly applicable to monochromatic radiation (i.e., radiation consisting of a single wavelength). Solutions for the entire PAR waveband may be obtained by dividing the waveband into narrow spectral intervals, solving the radiative transfer equation for each wavelength, and then adding the contributions of the wavelength intervals. Thus, RTT deals with spectral radiance as the most detailed descriptor of the radiation field. (Spectral) radiance  $R(\vec{\Omega})$ , sometimes erroneously termed radiation intensity, is defined as the radiative energy arriving from a given direction crossing a small (imaginary) surface element per unit solid angle per unit surface area. The SI unit of radiance is  $\text{W m}^{-2} \text{sr}^{-1}$  (Watts per square meter per steradian). Evidently,  $R(\vec{\Omega})$  is a function of the direction  $\vec{\Omega}$ , and thus describes the angular distribution of the radiation field. When dealing with the spectral characteristics of radiation, spectral radiance, or radiance per unit wavelength interval, is used. Similarly, when describing PAR, the PAR radiance  $R_{\text{PAR}}(\vec{\Omega})$  is used, or radiance carried by photons with wavelengths in the PAR waveband. The mathematical formulation of the theory does not depend on the wavelength interval and is the same for  $R(\vec{\Omega})$  and  $R_{\text{PAR}}(\vec{\Omega})$ . Similarly, the units of PAR radiance are those of the radiance  $R(\vec{\Omega})$ .

Integrating radiance over the hemisphere (corresponding to a solid angle of  $2\pi$ ) using cosine as the weighing function yields irradiance:

$$I(\vec{\Omega}) = \int_{2\pi} R(\vec{\Omega}') \cos(\widehat{\vec{\Omega}', \vec{\Omega}}) d\Omega', \quad (4)$$

where  $\widehat{\vec{\Omega}', \vec{\Omega}}$  is the angle between the directions  $\vec{\Omega}$  and  $\vec{\Omega}'$ . Irradiance  $I$  equals the amount of radiative energy carried through a unit area of the surface. A similar equation may be written to relate the PAR radiance  $R_{\text{PAR}}$  and the PAR irradiance  $I_{\text{PAR}}$ . From Eq. 4, it is evident that the irradiance  $I_{\text{PAR}}$  is a function of a directional variable  $\vec{\Omega}$  describing the normal of the surface: for any given surface, the amount of radiative energy it receives depends on its orientation. Commonly, irradiance is measured on a horizontal surface. When measuring downward-directed flux arriving from the upper hemisphere (i.e., incident flux),  $\vec{\Omega}$

points downward; when measuring reflected radiation,  $\vec{\Omega}$  points upward.

Radiation flux, or the amount of radiation crossing a surface in unit time, is calculated by dividing the surface into small surface elements, finding the flux density (i.e., irradiance) for each element, and finally adding the contributions of the surface elements. In mathematical terms, the summation is performed as an integration. In this way, a quantitative measure of radiation flow can be obtained through any (possibly imaginary) surface regardless of its shape and orientation. When measuring PAR flux for estimating photosynthesis (in either quantum or energy units), the surface should be that of a plant leaf. To find PPF on an arbitrarily inclined leaf surface, the angular distribution of the radiation field quantified by the radiance  $R_{\text{PAR}}(\vec{\Omega})$  has to be known. However, it is impractical, if not impossible, to measure the angular variation of radiance for each point inside a complex vegetation canopy. Under natural conditions, PAR arrives from the upper hemisphere only and to estimate the energy received, intercepted, and absorbed by a canopy, it is sufficient to measure fluxes on horizontal surfaces. This is in good accordance with the common practice of using the terms “irradiance” or “PPFD”: unless specified otherwise, fluxes are measured using a horizontal (and leveled) sensor. However, for the sake of clarity, it is advisable to always specify the directionality of the radiation receiving surface when describing flux measurements.

## Techniques for Measuring PAR

### Sensors Used for PAR Measurements

The actual sensors used to measure PAR vary in construction and the principle behind the radiation-to-voltage conversion. Two broad classes may be defined, corresponding roughly to instruments for measuring the two quantities defined in the previous section, PPF and PAR irradiance. Accordingly, two ideal PAR sensors may be defined: the ideal PAR quantum sensor, designed to measure PPF, and the ideal PAR energy sensor, to measure PAR irradiance. The definition of an ideal sensor is not based on its construction quality or working principle, but on its spectral response function  $\varepsilon(\lambda)$ . This function, similar to the



photosynthetic action spectrum, describes the output of the instrument when illuminated by a monochromatic radiation source with wavelength  $\lambda$ . To obtain the response of the sensor to any natural radiation source, the spectral response function has to be integrated over the spectral sensor's sensitivity range:

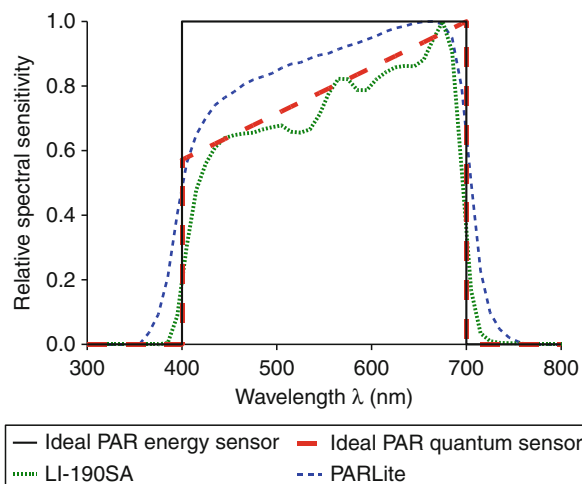
$$M_{\text{PAR}} = \int_{\lambda_{\text{min}}}^{\lambda_{\text{max}}} \varepsilon(\lambda) I_{\lambda}(\lambda) d\lambda, \quad (5)$$

where  $M_{\text{PAR}}$  is the sensor reading,  $\lambda_{\text{min}}$  and  $\lambda_{\text{max}}$  define the spectral interval where the sensitivity function is nonzero, and  $I_{\lambda}$  is the spectral irradiance. The sensor reading  $M_{\text{PAR}}$  is usually obtained in electric units: voltage or current produced by the sensor. While  $M_{\text{PAR}}$  is not directly usable for characterization of the radiation field, it is assumed to be proportional to the radiometric quantity of interest. The coefficient of proportionality, or calibration coefficient, is discussed in the next section.

The spectral sensitivity function for an ideal PAR energy sensor is constant with wavelength,  $\varepsilon_I(\lambda) = \text{const}$  inside the PAR waveband. The ideal PAR quantum sensor measures the number of incident photons independent of their wavelengths. To achieve this, Planck's law prescribes that the spectral sensitivity function of an ideal PAR quantum sensor has to be proportional to wavelength,  $\varepsilon_Q(\lambda) \sim \lambda$ , in the spectral interval of 400–700 nm. Outside the PAR waveband,  $\varepsilon_Q \equiv \varepsilon_I \equiv 0$ . The spectral sensitivity functions of the two ideal sensors are presented in Fig. 2 together with the response functions of two commercially available sensors.

Real PAR quantum sensors are usually photovoltaic sensors based on the photoelectric effect. Use of the photoelectric effect makes the response to the number of photons, regardless of their wavelengths, almost linear. This also makes PAR quantum sensors very responsive: they respond to changes in PPFD almost instantly and the upper limit on temporal sampling frequency is determined by the timescale of natural PAR changes, not the technical capabilities of the sensor.

Common photovoltaic sensors are, in principle, photodiodes working in photo-galvanic regime. A complete hemispheric field of view is achieved by



**Photosynthetically Active Radiation: Measurement and Modeling. Figure 2**

The relative spectral sensitivity functions (normalized so that the maximum value of each curve equals unity) of two ideal sensors for measuring PAR irradiance (Ideal PAR energy sensor) and PPFD (Ideal PAR quantum sensor) together with the curves for two real sensors (LI-COR LI-190SA and Kipp and Zonen PARLite)

placing a diffuser, a carefully shaped piece of diffusely transparent material, in front of the receiving element (the diode). The spectrally nonselective nature of the diffuser material in the PAR waveband makes the receiving surface look white. A suitable filter blocks out wavelengths outside the PAR region. Choice of the filter, along with the physical design of the instrument, brings the spectral response curve closer to that of an ideal sensor. The most widely used quantum sensor is the LI-190SA by LI-COR, Inc. [17, 18] which consists of a silicon photodiode covered by a visible band-pass interference filter and a colored glass filter. Since its introduction, use of quantum sensors to measure PAR has been expanding rapidly [19–24]. Currently, PAR sensors with silicon photodiodes are manufactured by several companies (e.g., PARLite by Kipp and Zonen, E90 Quantum sensor of Jauntering International Corp, SAT-LANTIC PAR sensor for underwater measurements) and form the most commonly used PAR sensor class.

Recently, GaAsP photodiodes have become available for use in PAR sensors (e.g., QSP-2100 by

Biospherical Instruments Inc., JYP 1000 by SDEC France). These sensors are inexpensive because the spectral sensitivity curve of a GaAsP photodiode is close to that of an ideal PAR quantum sensor. Wavelengths below 400 nm may be cut off by choosing a suitable material for the diffuser (usually polyacrylite), and special correction filters are not needed [25].

The second broad class of PAR sensors, PAR energy sensors, includes mostly thermoelectric instruments. These instruments are designed to measure PAR irradiance with a constant sensitivity function using a black receiving surface which is heated by incident radiation. Using a calibration coefficient, the temperature reading of the receiving surface is converted into irradiance. The receiving surface is covered by a glass filter to block photons with wavelengths outside the PAR waveband. Compared with quantum sensors, thermoelectric instruments are technically more complicated and expensive. However, such instruments can be constructed with the same bodies as standard short-wave solar radiation measurement devices, pyranometers and pyrhemometers, making the measurements robust and repeatable. The first hemispheric measurements of global, diffuse, and reflected PAR were made using pyranometers covered with hemispherical glass filters [14, 26–33] and measurements of direct solar PAR with pyrhemometers covered with flat glass filters [34–36]. A thermoelectric device for measuring submarine PAR was also constructed from a thermopile coated with Parsons' black lacquer and covered by a glass filter [37]. Compared to quantum receivers, thermoelectric instruments exhibit large inertia: changes in the temperature of the receiving surface follow the changes in irradiance after a substantial time delay. The time constant (the time it takes for the output signal to decrease by  $e \simeq 2.72$  times after incident radiation is completely blocked) of common thermoelectric instruments is about 10 s. Although such timescales are reasonable when measuring incident PAR in the open (affected mainly by solar elevation and atmospheric transmission), variations in radiation fluxes inside a plant canopy happen on much shorter timescales.

An interesting novel idea is to use light emitting diodes (LEDs) in photo-galvanic regime as radiation sensors. A combination of blue and red LEDs provides

an acceptable approximation of the PAR spectral curve. An inexpensive sensor consisting of blue SiC and red GaP or AlGaAs LEDs exhibited good correlation with the LI-COR LI-190SA quantum sensor when measuring global PAR [38].

The most complete way to describe radiation in the PAR waveband is to measure its spectral composition. Unfortunately, the instruments for spectral radiation measurements, spectroradiometers (more commonly called just spectrometers), have been expensive and not well suited for field measurement or long-time automatic monitoring. In recent years, developments in affordable photodiode array technology have made the construction of spectroradiometers with few or no moving parts possible. These lightweight field instruments typically measure radiation between 350 and 1,050 nm with a sampling interval of a few nanometers. However, judging from the relatively small number of published results, the simplicity, robustness, and low price of quantum sensors outweigh the increased amount of data and the higher price tag produced by a spectroradiometer. In many common applications in agriculture, horticulture, or monitoring of photosynthetic productivity, monitoring of the amount of available PAR, rather than its spectral composition, is sufficient. Nevertheless, advances in technology indicate that in the decades to come, radiometry will be shifting from broadband sensors toward spectral instruments.

Most sensors described above are intended to measure global PAR: they have been designed to integrate the radiation arriving from all directions in a hemisphere and output radiation flux density. Such sensors are also called cosine receivers after the weighing function used in the mathematical formulation of the integration formula (Eq. 4). Thus, the field of view (FOV) of a cosine receiver is  $2\pi$ , the solid angle corresponding to a hemisphere. Sometimes, the FOV of an instrument is restricted to receive photons coming from a single direction, for example, the sun. Alternatively, to measure only diffuse sky radiation, sensors may be equipped with a shadow band blocking the diurnal path of the sun in the sky, or a tracking shade disc (i.e., a small disc mounted on an arm activated by a mechanical device used to keep scientific instruments directed toward the sun). To measure the direct solar

component of PAR, that is, the flux density of PAR not scattered by the atmosphere, a pyrliometer may be equipped with glass filters; alternatively, a common hemispheric PAR sensor may be fitted with a view-limiting tube analogous to that of a pyrliometer [39]. Unfortunately, no PAR quantum sensors specially designed to measure direct radiation are commercially available. A narrow FOV is also used to study the directional properties of radiation field: directional reflectance or directional distribution of incident radiation [40, 41]. Additionally, instruments to measure radiation arriving from all directions (corresponding to a solid angle of  $4\pi$ ) have been designed. Such instruments measure a quantity called radiation fluence rate and they are more commonly used in aquatic environments (see section “Description of PAR in Water”).

To quantify the enormous variability of the radiation field inside and below a plant canopy, single sensors do not suffice. Elaborate systems can be combined with consumer equipment to obtain the best results. The measurement systems used in plant canopies are briefly discussed in section “Instruments for Measuring fAPAR”. Although only a few of these devices include the PAR sensors described above, the implicit physical principles of radiometry in these instruments, and thus also the inherent limitations and potential errors, are exactly the same.

### Calibration and Spectral Corrections

Direct comparison of two PAR sensors is not a simple task. The reading of a PAR sensor can be predicted from the reading of another sensor if the spectral sensitivity functions of both sensors are known as well as the spectral composition of incident radiation. Although most producers of PAR sensors provide the spectral response functions for their instruments, the spectral composition of incident radiation is generally unknown: the relatively stable spectral composition of extraterrestrial solar PAR is heavily altered when passing through the atmosphere. Inside a plant canopy, the spectral composition of PAR is further distorted by the removal of photons at blue and red wavelengths, where the absorbance of plant leaves is the highest. Thus, care must be taken when comparing the

numerical outputs of different sensors in radiation absorption measurements as well as during calibration.

To calibrate a PAR sensor, one needs to measure its output in a controlled experimental situation (e.g., using a calibration lamp) where the value of the measured quantity is known. A calibration coefficient is the ratio of the actual value of the measurable quantity (e.g.,  $I_{\text{PAR}}$ ) to the instrument reading ( $M_{\text{PAR}}$ ):

$$\mu_I = \frac{I_{\text{PAR}}}{M_{\text{PAR}}} = \frac{\int_{400}^{700} I_{\lambda, \text{LAMP}}(\lambda) d\lambda}{\int_{\lambda_{\text{min}}}^{\lambda_{\text{max}}} \varepsilon(\lambda) I_{\lambda, \text{LAMP}}(\lambda) d\lambda}. \quad (6)$$

Two calibration coefficients are usually provided for PAR sensors: one to convert the sensor’s reading into energy units ( $\text{W m}^{-2}$ ) defined by Eq. 6 and one for quantum units ( $\mu\text{mol m}^{-2} \text{s}^{-1}$ ). The coefficient for quantum units is defined similarly to that for energy units,  $\mu_Q = Q_{\text{PAR}}/M_{\text{PAR}}$ .

Most manufacturers calibrate the sensors in laboratory using standard lamps, which is indicated by using the subscript LAMP for the spectral irradiance in Eq. 6,  $I_{\lambda, \text{LAMP}}$ . Thus, the manufacturers can control (within a given measurement uncertainty of a few percent determined by the calibration of the lamp) both the energy content and the spectral composition of incident radiation. If the spectral sensitivity of the actual sensor being calibrated deviates from that of the perfect sensor, as it invariably does, the calibration coefficient depends on the spectral composition of incident radiation. Therefore, the laboratory-derived calibration is only directly valid for irradiation by a calibration lamp and does not hold exactly under field conditions.

The errors arising from the mismatch of the predicted and actual field conditions are usually called spectral errors. Spectral errors depend on the spectral sensitivity of the sensor, the spectral composition of incident irradiance, and the spectral composition of radiation used to calibrate the sensor. Mathematically, the spectral error in PAR irradiance measurements can be written as

$$\beta_I = \frac{\mu_I \int_{\lambda_{\text{min}}}^{\lambda_{\text{max}}} \varepsilon(\lambda) I_{\lambda}(\lambda) d\lambda}{\int_{400}^{700} I_{\lambda}(\lambda) d\lambda}. \quad (7)$$

After substituting the expression for  $\mu_I$  from Eq. 6 into Eq. 7, it becomes clear that spectral errors

disappear if (1) the sensor has a response function identical to that of the ideal sensor,  $\varepsilon(\lambda) \equiv \varepsilon_I(\lambda)$ , or (2) the irradiance conditions match the calibration lamp spectrum,  $I_\lambda(\lambda) \equiv I_{\lambda, \text{LAMP}}(\lambda)$  in the spectral interval from  $\lambda_{\text{min}}$  to  $\lambda_{\text{max}}$ . The magnitude of the actual spectral error varies with sensor type. While it is reasonably small for the LI-190SA sensor, usually less than 1% for natural irradiance conditions, many other sensors exhibit considerably larger errors, especially under artificial illumination [11].

If the spectral composition of incident radiation or, more specifically, the difference between the spectral composition of radiation occurring during field measurements and during calibration, spectral error can be eliminated by using a spectral correction. It is evident that in the presence of the spectral error, multiplying the measurement result by the inverse of  $\beta_I$  (Eq. 7) would compensate completely for the differences in the spectra of incident radiation. Thus, if the actual spectral irradiance  $I_\lambda(\lambda)$  is known, a spectral correction can be easily calculated. However, since  $I_\lambda(\lambda)$  is usually not available, a general value characterizing the illumination conditions (clear or cloudy sky, different artificial light sources),  $I_{\lambda, \text{EST}}(\lambda)$ , is used instead. Thus, the spectral correction factor is calculated as the reciprocal of  $\beta_I$  after replacing  $I_\lambda(\lambda)$  by  $I_{\lambda, \text{EST}}(\lambda)$  in Eq. 7.

Therefore, when taking a measurement, the instrument reading is first multiplied by the calibration coefficient ( $\mu_I$  or  $\mu_Q$ , calculated individually for each sensor) and, optionally, by a spectral correction (calculated for a whole instrument class or model). As only the average spectral irradiance distribution for typical conditions is known, it is often preferable to ignore spectral corrections.

Another possibility to calibrate the sensors is to compare them with a reference sensor (or a spectroradiometer) with a reliable calibration by the manufacturer. When performing calibration under irradiance conditions reasonably close to those occurring under true measurement situations, spectral corrections are not required. For example, field calibrations of radiation measuring instruments are standard for the Baseline Surface Radiation Network (BSRN, [42]), an international network for global measurements of solar and atmospheric radiation at the highest available accuracy. Additionally, all PAR sensors continuously exposed to outdoor conditions should be

checked regularly against well-maintained reference instruments. The sensitivity of sensors is apt to change due to the aging of the diffuser and filters. Such aging is usually also documented in the instrument manual. When a frequently and reliably calibrated instrument is not available, it is strongly recommended to have a reference instrument stored under controlled conditions for periodical comparisons with the operating sensors.

### Measurement Errors

As with all measurements, errors are inevitable and arise from (1) the impossibility of controlling all the physical processes that determine the measurement result, (2) the non-perfect construction of the measuring apparatus, and (3) the spectral composition of incident radiation. Since the last error source (spectral errors) was covered in section “[Calibration and Spectral Corrections](#)”, only the first two categories are briefly discussed here.

Measurement errors can be reduced by carefully following the instructions for performing the measurements (usually provided by the manufacturer), using proper installment and maintenance procedures (e.g., checking for the directionality and leveling of the instrument), checking the performance of the instrument regularly, and accounting for material degradation and changes in operating environment (ambient temperature, irradiance conditions, humidity, etc.). Flux measurements with a hemispherically integrating sensor suffer from directionality effects: the sensor is not equally sensitive to radiation arriving from different directions. While manufacturing imperfections or physical damage may cause an instrument to have random sensitivity fluctuations with the azimuth angle of an incident beam, sensitivity to polar angle (or zenith angle for a leveled sensor looking upward) of the radiation source is usually more systematic. The dependence of sensitivity with the polar angle is called the cosine response of the sensor and the corresponding correction a cosine correction. The cosine response characteristics of several sensors designed for irradiance measurements, including two LI-COR 190 PAR sensors, are given in [43]. Cosine effects, together with leveling inaccuracies, are especially influential when a strong directional radiation

source is present, such as the direct solar radiation beam on a clear day. All these errors, some systematic and some random, can add to the spectral errors affecting instruments designed for measuring radiation in a spectral waveband as discussed in the previous section.

The official relative uncertainty of PAR instruments claimed by manufacturers, about 5%, can only be achieved under optimal conditions. During routine measurements, even if performed by trained specialists, the uncertainty can be considerably larger. For example, [44] gives an estimate of 10% uncertainty for PAR measurements in the FLUXNET network; a comparison performed by BSRN found significant systematic differences between different PAR sensor models and up to 20% spread within the group of 11 tested LI-190SA sensors [45].

## PAR in Various Environments

### PAR Below the Atmosphere

Without the influence of the atmosphere, the PAR irradiance would be determined by the solar spectrum and geometric conditions like the slightly varying distance from the earth to the sun, local solar elevation, and topographic shadowing. The spectral composition of radiation would be constant to the accuracy of the multiple scattering contribution of non-flat topography (illuminated slopes of mountains and valleys). Such direct topographic effects, although significant in shadow areas, are usually small when direct solar radiation is present and will be ignored hereafter. Under natural conditions, the amount and spectral composition of radiation in the PAR spectral band, in addition to the distance from the sun to the earth and solar elevation angle, is mainly determined by the presence of clouds, the amount and optical properties of aerosols, and, to a lesser extent, the chemical composition of the atmosphere.

Due to its universal nature, radiative transfer theory (RTT, section “[Fundamentals of Radiation Transfer Theory](#)”) can be (and also has been) applied to predict the irradiation conditions under all possible atmospheric conditions. The actual precision of prediction is limited by the availability of input data and computer power. Models based on RTT can be used to calculate accurately the spectral irradiance for the different wavelengths comprising PAR, with subsequent

integration to obtain  $I_{\text{PAR}}$ . For many practical purposes, however, simpler models applicable to longer timescales (hours, days, growing seasons) are sought and thus different broad-band models or physically based parametrizations are often used. For clear skies, the accuracy of the best broadband models is comparable to that of routine irradiance measurements in existing networks [46].

Models developed for predicting the behavior of sunlight in the atmosphere deal not only with PAR but with the whole shortwave spectral region. The shortwave spectral region is loosely defined as the range of wavelengths containing the bulk of the solar spectrum (magnitude wise), usually between 300 and 4,000 nm. The simplest case, global shortwave irradiance under a cloudless atmosphere, can be very accurately predicted when the following parameters are known (REST2 model, [46]): solar zenith angle, Ångström turbidity coefficient (i.e., aerosol optical depth at 1,000 nm), Ångström wavelength exponents, aerosol single-scattering albedo, air pressure, amounts of precipitable water and ozone, and ground albedo. These parameters allow to calculate the irradiance in two separate wavebands, PAR and short-wave infrared. To predict only PAR irradiance, a few parameters less are required, since ozone and water vapor have little influence on PAR.

Because not all of the listed atmospheric parameters are readily available, models based on easily measurable radiation field characteristics have also been developed [24, 47, 48]. Usually, parametrizations are based on approximately four parameters. While one parameter is always solar elevation, others describe the state of the atmosphere and can be either obtained from radiation measurements (e.g., ratio of diffuse to direct shortwave irradiance) [49] or routine meteorological data (e.g., dew point temperature). The variables explaining the majority of variance in PAR availability and the spectral quality of PAR ( $Q_{\text{PAR}}/I_{\text{PAR}}$  ratio) include solar elevation and a parameter to describe the turbidity of the atmosphere (e.g., a sky clearness parameter or the Ångström turbidity coefficient) [39, 47, 48, 50, 51].

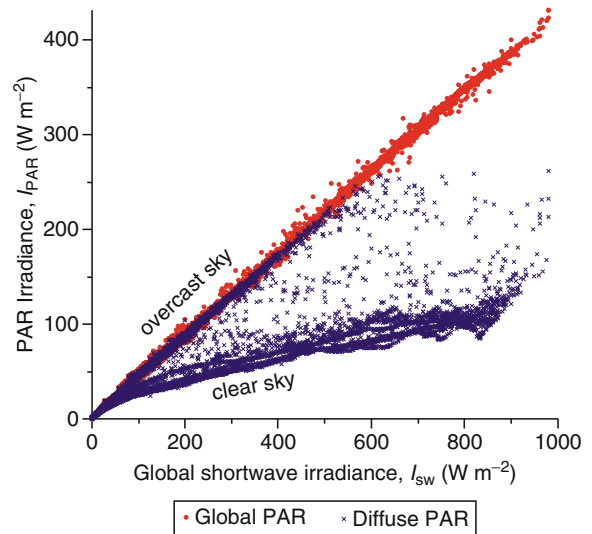
Two downwelling PAR field components can be distinguished under a clear sky: the quasi-parallel direct beam arriving from the direction of the sun (with PAR irradiance on a horizontal surface  $I_{\text{PAR,dir}}$ ) and the diffuse sky radiation arriving from all upward directions not blocked by topography ( $I_{\text{PAR,diff}}$ ).

The sum of the two components is called global PAR,  $I_{\text{PAR}} = I_{\text{PAR,dir}} + I_{\text{PAR,diff}}$ . Depending on aerosol load and solar elevation, the ratio of diffuse PAR to global PAR irradiance on a horizontal surface ranges between 20% and 40% [52]. The presence of clouds decreases the global PAR irradiance usually by up to 80% [41] and the contribution of diffuse PAR irradiance may take any value between that corresponding to a clear sky and 100%. If the cloud cover is broken, the existence of the direct beam depends on the locations of gaps between clouds. Thus, two temporally and spatially variable phenomena have great influence on the amount of diffuse PAR: aerosol loading and clouds. Although the effect of aerosols may be dominating in places with low average cloud cover, it is expected that the contribution of clouds as the source of variations in diffuse PAR is larger for most locations.

The fraction of PAR in global shortwave irradiance  $I_{\text{PAR}}/I_{\text{SW}}$  varies little and is usually between 40% and 50% [53, 54]; values above 50% occur under very low sun, thick cloud cover, or rain [14]. As an example, measurements at Tõravere actinometric station (Estonia) are presented in Fig. 3 for variable cloud conditions during June 2009. The global PAR irradiance can be relatively reliably predicted from global shortwave irradiance,  $I_{\text{PAR}} = 0.43 I_{\text{SW}}$ . The contribution of diffuse PAR irradiance  $I_{\text{PAR,diff}}$  to global shortwave irradiance, on the other hand, was more variable (Fig. 3). Under completely overcast skies, the diffuse PAR irradiance  $I_{\text{PAR,diff}}$  equals the global PAR irradiance  $I_{\text{PAR}}$ , which, as usual, contributed about 43% of global shortwave irradiance. Under clear skies,  $I_{\text{PAR,diff}}$  is significantly smaller than  $I_{\text{PAR}}$ : in Fig. 3 the data points corresponding to clear sky form the lower cluster. Broken cloud cover conditions are represented by  $I_{\text{PAR,diff}}$  values between the two extremes when plotted against  $I_{\text{SW}}$ .

Some variation in  $I_{\text{PAR}}/I_{\text{SW}}$  with elevation above sea level is expected, but this variation is difficult to detect [52]. However, using measurement sites at 550, 900, and 1,500 m above sea level, an increasing trend was noted with altitude, of 3.6% per km for hourly values of  $Q_{\text{PAR}}/I_{\text{SW}}$  under clear skies [55]. An inverse trend was found for hourly  $Q_{\text{PAR}}/I_{\text{SW}}$  under cloudy weather conditions:  $Q_{\text{PAR}}/I_{\text{SW}}$  decreased at a rate of 1.8% per km.

The spectral composition of global PAR is relatively stable [56]. It is reflected in the near-constant value of



**Photosynthetically Active Radiation: Measurement and Modeling. Figure 3**

Global and diffuse PAR irradiance as functions of global shortwave irradiance in Tõravere, Estonia, in June 2009. Sky condition varied from clear to completely overcast. The labels “clear sky” and “overcast sky” indicate the characteristic values of diffuse PAR irradiance for the two atmospheric conditions

the ratio of PPFD to PAR irradiance,  $Q_{\text{PAR}}/I_{\text{PAR}}$  (Eq. 3). The classical value of  $Q_{\text{PAR}}/I_{\text{PAR}} = 4.57 \mu\text{mol s}^{-1} \text{W}^{-1}$  for global PAR proposed by McCree [2] has been verified by several later studies. For example, [57] reported that while 1-min average of  $Q_{\text{PAR}}/I_{\text{PAR}}$  varied from 4.23 to 4.68  $\mu\text{mol s}^{-1} \text{W}^{-1}$ , 1-h averages were relatively insensitive to atmospheric composition with  $Q_{\text{PAR}}/I_{\text{PAR}} = 4.56 \mu\text{mol s}^{-1} \text{W}^{-1}$  for global PAR.

For the diffuse radiation field component, the ratio depends on atmospheric conditions. Under a blue sky, an average value of  $Q_{\text{PAR,diff}}/I_{\text{PAR,diff}} = 4.28 \mu\text{mol s}^{-1} \text{W}^{-1}$  was reported [52] along with the observation that the ratio increases with aerosol load. In the presence of clouds,  $Q_{\text{PAR,diff}}/I_{\text{PAR,diff}}$  increases with increasing cloud cover from 4.24  $\mu\text{mol s}^{-1} \text{W}^{-1}$  (a value characteristic of blue sky) to the constant value for global radiation,  $Q_{\text{PAR}}/I_{\text{PAR}} = 4.57 \mu\text{mol s}^{-1} \text{W}^{-1}$  under an overcast sky [57]. The value of  $Q_{\text{PAR}}/I_{\text{PAR}}$  (or  $Q_{\text{PAR,diff}}/I_{\text{PAR,diff}}$ ) describes the color of light: the smaller the ratio, the bluer the light looks to the human eye.

The angular distribution of PAR radiance can be approximated using models applicable to visible

light [16]. Additionally, several approximations exist for predicting the angular distribution of shortwave radiation. These models have been parametrized for use in the PAR waveband [40, 41] for different atmospheric conditions ranging from completely clear to overcast sky. The angular models describe sky radiance relative to the nadir direction and cannot be generally used to describe global PAR irradiance or PPFD.

A final remark on the spectral quality of PAR at the bottom of the atmosphere can be made based on the spectral composition of extraterrestrial solar radiation. The  $I_{\text{PAR}}/I_{\text{SW}}$  ratio outside the atmosphere based on the solar constant of  $I_{\text{SW}} = 1367 \text{ W m}^{-2}$  equals 38.8% [51]. Using  $Q_{\text{PAR}} = 2426 \mu\text{mol s}^{-1} \text{ m}^{-2}$  for the extraterrestrial irradiance on a surface perpendicular to sunrays [48], we obtain that the  $Q_{\text{PAR}}/I_{\text{PAR}}$  ratio outside the atmosphere equals  $4.57 \mu\text{mol s}^{-1} \text{ W}^{-1}$  – exactly that proposed by McCree [2]. While it may be concluded that the atmosphere has little effect on the spectral quality of PAR, the exact coincidence is most likely due to chance: an accuracy of two decimals is clearly beyond the uncertainties inherent in radiation measurements.

### Description of PAR in Water

The waveband of radiation allowing phytoplankton to carry out photosynthesis (i.e., PAR) corresponds approximately to the same spectral band of electromagnetic radiation that penetrates into water. Pure water absorbs strongly in the ultraviolet ( $\lambda < 400 \text{ nm}$ ) and near-infrared ( $\lambda > 700 \text{ nm}$ ) spectral regions [58]. Otherwise, the underwater light field is determined by the incident irradiance (see section “PAR Below the Atmosphere”), the state and composition of the water body, and the optical properties of its bottom.

The spectrum of solar radiation penetrating a water body changes drastically as its irradiance diminishes with depth. While the scattering of radiation is commonly rather insensitive to wavelength in the PAR waveband, absorption by different components has a very strong spectral effect. The components having an optical effect are dissolved organic substances (also known as yellow substance), different species of phytoplankton, and inert particulate matter (Fig. 4). Since the concentration of these is highly variable, the

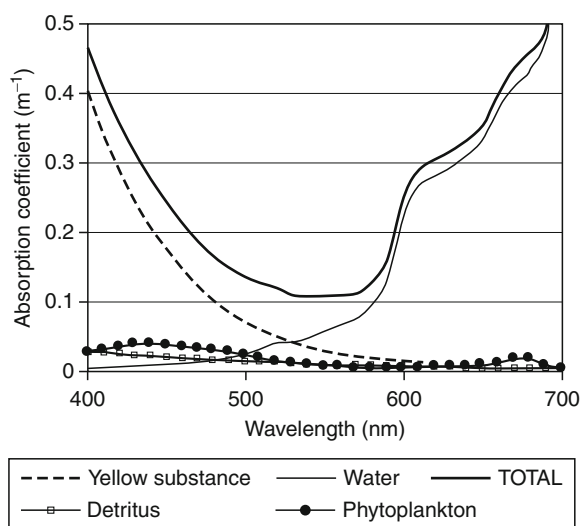
spectral distribution of underwater irradiance in the PAR waveband can change rapidly.

In a vertically homogeneous water body, the value of the downwelling spectral irradiance  $I_{\lambda}$  diminishes approximately exponentially with depth  $z$ , that is, Beer’s law holds (see also Eq. 8):

$$I_{\lambda}(z) = I_{\lambda}(0) \exp \left[ - \int_0^z K_{\text{diff}}(\lambda, z') dz' \right]$$

where  $K_{\text{diff}}(\lambda, z)$  is the diffuse attenuation coefficient, a parameter often used to describe the optical properties of natural water bodies [62–64].

In addition to the spectral PAR irradiance  $I_{\lambda}$  and PPFD, a quantity called spectral fluence rate, or spectral spherical irradiance, is sometimes used to describe the amount of radiation in water. It is defined as the total amount of photons incident in unit time interval from all directions on a small sphere, divided by the



### Photosynthetically Active Radiation: Measurement and Modeling. Figure 4

Decomposing the absorption spectrum of a water sample. Spectra of the absorption coefficients corresponding to pure fresh water [59];  $1.0 \times 10^{-3} \text{ mg m}^{-3}$  yellow substance (specific absorption coefficient at 380 nm  $0.565 \text{ L mg}^{-1} \text{ nm}^{-1}$ );  $1.0 \text{ mg m}^{-3}$  chlorophyll-a (phytoplankton, [60]); and detritus (from the measured data of [61]). The total absorption of the water sample is plotted with a bold line

cross-sectional area of the sphere. Analogously to the PAR irradiance  $I_{\text{PAR}}$ , the PAR fluence rate may be defined by integrating over the PAR waveband. As both fluence rate and irradiance describe the amount of PAR in water, Beer's law can (with a different attenuation coefficient) also be applied to describe the change in spectral fluence rate with depth.

Beer's law is valid only for the spectral irradiance  $I(\lambda)$ , that is, the irradiance in a narrow spectral interval around wavelength  $\lambda$ . Many authors have shown that the exponential law fails when a single value of  $K_{\text{diff}}$  is applied over the whole PAR waveband. The reason for this failure lies in the change of spectral composition of PAR with depth. This can be illustrated using the wavelength corresponding to maximum penetration,  $\lambda^*$ , and the  $Q_{\text{PAR}}/I_{\text{PAR}}$  ratio. In clear oceanic waters,  $\lambda^*$  corresponds to the maximum in the extraterrestrial solar spectrum ( $\sim 460$  nm). When increasing the amount of optically active substances in water,  $\lambda^*$  is shifted toward larger values, as shown in Table 1, and can be even larger than 700 nm in brownish boreal lakes [65].

Photosynthesis in water takes place mainly in the so-called euphotic layer near the surface. At the bottom of the euphotic layer, the downward PAR irradiance has decreased to 1% of its value just below the surface [66]. In clear oceanic waters, the thickness of the euphotic layer can be of the order of a hundred meters. As a contrast, in turbid lakes, the layer may be only half a meter thick. Ice cover, and especially ice covered with snow, may substantially decrease the amount of PAR in water to a level not sufficient for even the minimum amount of photosynthetic activity, thus creating anoxic conditions [67].

Similarly to other environments, the PAR irradiance in water can be given in energy units

( $I_{\text{PAR}}$ ,  $\text{W m}^{-2}$ ) or quantum units ( $Q_{\text{PAR}}$ ,  $\mu\text{mol s}^{-1} \text{m}^{-2}$ ). The quanta-to-energy ratio  $Q_{\text{PAR}}/I_{\text{PAR}}$  ( $\mu\text{mol s}^{-1} \text{W}^{-1}$ ) changes with the variation of the spectral distribution of irradiance. Above water,  $Q_{\text{PAR}}/I_{\text{PAR}}$  is practically constant with an average value  $Q_{\text{PAR}}/I_{\text{PAR}} = 4.57 \mu\text{mol s}^{-1} \text{W}^{-1}$  over a wide range of conditions (see section "PAR Below the Atmosphere"). In clear oceanic water,  $Q_{\text{PAR}}/I_{\text{PAR}}$  decreases with depth. As a contrast, in turbid coastal waters and lakes, it increases with depth and approaches an asymptotic value. Thus, sufficiently deep below the surface of turbid waters, the spectral distribution of PAR, but not the value of PAR irradiance, can be considered almost constant. The average value of  $Q_{\text{PAR}}/I_{\text{PAR}}$  there has been estimated at  $4.15 \pm 0.40 \mu\text{mol s}^{-1} \text{W}^{-1}$  [68]; a value of  $Q_{\text{PAR}}/I_{\text{PAR}} = 4.45 \pm 0.48 \mu\text{mol s}^{-1} \text{W}^{-1}$  has been suggested for Norwegian coastal waters [69]. In lakes,  $Q_{\text{PAR}}/I_{\text{PAR}}$  varies from 4.72 to  $5.86 \mu\text{mol s}^{-1} \text{W}^{-1}$  [65]. Additionally, there is a strong linear correlation between  $Q_{\text{PAR}}/I_{\text{PAR}}$  and  $K_{\text{diff}}$  ( $r = 0.95$ ) and, in deeper waters,  $Q_{\text{PAR}}/I_{\text{PAR}}$  can be estimated using  $K_{\text{diff}}$  measurements in the surface layer [65].

### Measurement Stations and Networks

Unfortunately, no international network to measuring PAR currently exists. As can be seen from their documentation, large international radiation measurement networks like the Baseline Surface Radiation Network (BSRN <http://www.gewex.org/bsrn.html>) [70] have discussed the subject of PAR measurements, but standardized measurements have not started. The PAR irradiance (and also APAR) is recorded as a by-product in some networks specialized in other measurements. For example, the FLUXNET project (<http://daac.ornl.gov/FLUXNET/>) [71, 72], which is aimed at

**Photosynthetically Active Radiation: Measurement and Modeling.** Table 1 Wavelength of maximum penetration  $\lambda^*$ , ratio  $Q_{\text{PAR}}/I_{\text{PAR}}$ , and relative difference  $D$  of  $Q_{\text{PAR}}/I_{\text{PAR}}$  from its value above the surface for different water types as classified by [62]

Water type	I	II	III	1	3	5	7	9
$\lambda^*$ (nm)	465	480	505	530	540	547	565	582
$Q_{\text{PAR}}/I_{\text{PAR}}$ ( $\mu\text{mol s}^{-1} \text{W}^{-1}$ )	3.9	4.0	4.2	4.4	4.5	4.5	4.7	4.9
$D$ (%)	16	14	9	4	2	2	2	6

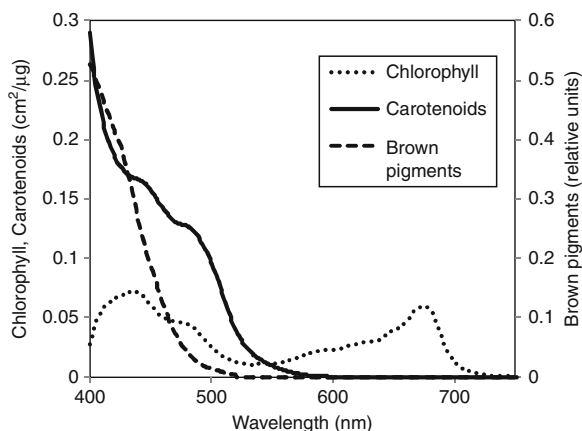


quantifying the exchanges of carbon dioxide, water vapor, and energy between the biosphere and atmosphere, has PAR data available for many sites. The solar radiation budget network SURFRAD (<http://www.srrb.noaa.gov/surfrad/>) [73] measures, among other variables, the incident PAR irradiance. A promising start is SolRad-Net (<http://solrad-net.gsfc.nasa.gov/>), a companion to the successful global AERONET aerosol network. However, the number of SolRad-Net sites where PAR is measured today is still very small. Routine monitoring of PAR and APAR as key factors in global photosynthetic productivity [74] has been proposed several times. Currently, the only global data sets available are those based on remote sensing data (see section “APAR and fAPAR from Satellite Observations”). Although remote sensing can provide excellent spatial coverage unachievable in any ground-based network, indirect remote retrievals should still be validated against direct measurements of the variable under investigation.

## PAR in Vegetation Canopies

### PAR Absorption by Leaves

**PAR Absorption by Leaf Pigments** Electromagnetic radiation in the PAR spectral domain is mainly absorbed by photosynthetic pigments in the leaf. Among these, chlorophylls a and b are the most important. They are found across a wide range of species, from algae to higher plants, and they participate in transforming radiation into energy, which is later stored as chemical bonds of carbohydrates. Chlorophylls are characterized by two absorption peaks at 450 and 670 nm corresponding to the blue and red color, respectively, explaining the green color of leaves (Fig. 5). Besides chlorophylls, green leaves contain also other pigments. Pigments such as carotenes and xanthophylls belonging to the carotenoid family associated with chlorophylls are known to improve radiation harvesting. They mainly absorb in the blue region with absorption peaks at 450 and 470 nm making them look orange or yellow. Additionally, carotenoids prevent oxidation of the photosynthetic system in case of excess incident radiation. Other pigments like anthocyanins absorb radiation in the PAR waveband with maximum absorption between 450 and 600 nm [77]. They protect the leaf against UV radiation by

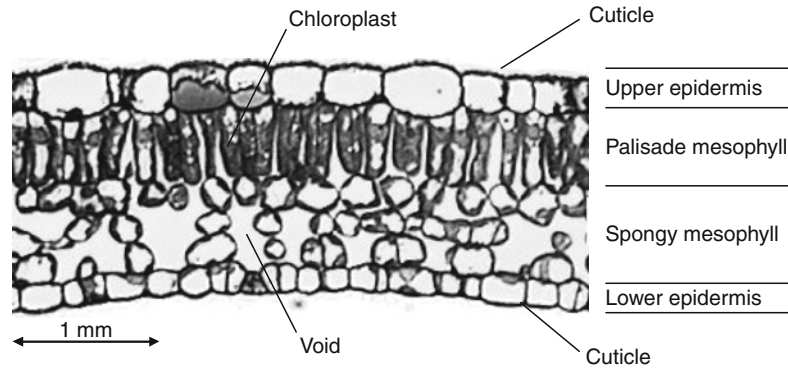


**Photosynthetically Active Radiation: Measurement and Modeling. Figure 5**

Specific absorption coefficients of chlorophyll (a + b), carotenoids and brown pigments from the PROSPECT model [75, 76]

preventing formation of free radicals and are responsible for the reddish colors of some leaves during the autumn. The rest of the biochemical leaf constituents responsible for the brown color (such as polyphenols, which develop during leaf senescence), although absorbing in the PAR waveband (Fig. 5), have no direct role in photosynthetic processes.

**The Role of Leaf Structure** The efficiency with which a leaf absorbs visible radiation depends not only on chlorophyll content per unit leaf area, but also on the specific mechanisms plants have developed to utilize radiation. Chlorophylls are concentrated in chloroplasts, mainly located in the palisade mesophyll consisting of tightly packed elongated cells just under the upper epidermis (Fig. 6). The tubular shape of palisade cells enhances the forward propagation of PAR, thus directing photons to the chloroplasts located at the bottom of the palisade. In a number of species, the epidermis cells act as lens focusing radiation on the chloroplasts thus increasing radiation absorption by the photosynthetic pigments [78]. The spongy mesophyll of higher plants contains little PAR-absorbing pigments. Instead, the numerous voids in this layer act as a mirror to scatter back a large fraction of the radiation transmitted through the palisade mesophyll, further improving absorption of radiation by chloroplasts.

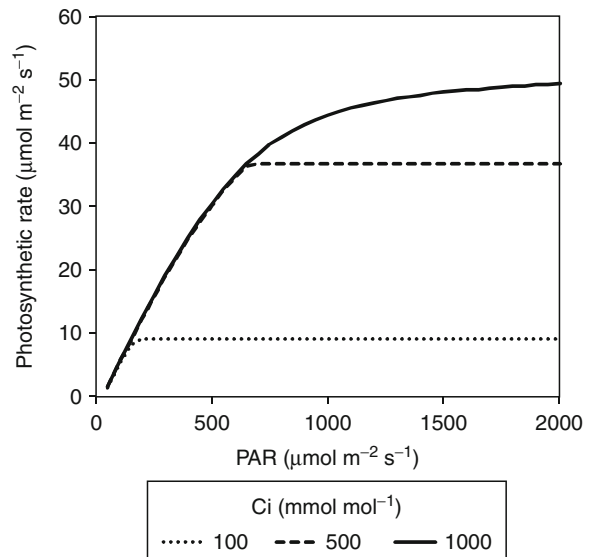


**Photosynthetically Active Radiation: Measurement and Modeling. Figure 6**  
Structure of a typical dicotyledon leaf (young maple)

When the capacity of the leaf to use the available PAR is exhausted by drought or nutrient stress, plants may use different strategies to minimize absorbed radiation. Besides changing the orientation of a leaf away from direct sunlight, adaptations have been developed at the upper epidermis level. The cuticle may turn to crystalline and the amount of hair on the leaf surface may increase, thus also increasing the leaf reflectivity and directing radiation away from chloroplasts.

**The Fate of Absorbed PAR in the Leaf** The energy carried by photons absorbed by the leaf is transformed into several types of energy. The dominant type is heat, accounting for more than 75% of the absorbed PAR energy. Therefore, only a maximum of 25% of APAR is left for photosynthesis. When photosynthesis is limited by temperature, water, or nutrient availability, the radiation use efficiency of a leaf decreases even further. Additionally, a part of the excess PAR energy absorbed by the pigments may be dissipated as fluorescence, that is, reradiated as photons with different, longer wavelengths.

Under optimal water, temperature, and nutrient conditions, the efficiency of photosynthesis at the leaf level is determined both by the absorbed PPFD and the  $\text{CO}_2$  concentration in the leaf [79]. Figure 7 shows that the photosynthetic rate increases almost linearly with incident PPFD up to some threshold value. After the threshold is reached, the rate of photosynthesis becomes constant, limited by the  $\text{CO}_2$  concentration in the leaf. This concentration in the leaf, in turn, is driven by the  $\text{CO}_2$  concentration in the atmosphere



**Photosynthetically Active Radiation: Measurement and Modeling. Figure 7**

Relation between incident PAR photon flux density and photosynthesis rate at different intercellular concentration of  $\text{CO}_2$  ( $C_i$ ), simulated using model of C3 photosynthesis [79]

and the stomatal conductance of the leaf, itself determined by the hydraulic status of the leaf and the plant as a whole [80]. The maximum photosynthetic rate, achieved at large PPFD values, also strongly depends on the fraction of the incident PAR absorbed by the leaf [81], which is determined mostly by leaf chlorophyll content.

Since chlorophyll absorbs the majority of incident PAR and uses only a small fraction of it for photosynthesis, the excess energy must be dissipated in a way that keeps the photosynthetic apparatus functional. For this purpose, plants have developed several photoprotection mechanisms. One of such mechanisms is based on xanthophyll pigments: the conversion of violaxanthin into zeaxanthin can remove the excess energy from chlorophyll and dissipate it as heat. When conditions become more favorable, this conversion is reversed, and zeaxanthin is changed back to violaxanthin. The status of the xanthophyll cycle may be used to evaluate the various stresses experienced by plants, since it affects the leaf optical properties, or leaf reflectance spectrum, in the 500–600 nm spectral region [82].

The second pathway for dissipating excess energy absorbed by photosynthetic pigments is fluorescence. When photosynthesis is limited by stress factors or when a leaf is exposed to too high irradiance, a small but measurable fraction of the excess energy is reemitted at a longer wavelength than that of absorption. The peak of chlorophyll fluorescence emission is in the blue-green (455 nm), red (685 nm), and far-red (735 nm) spectral regions [83]. The energy lost in this process amounts to a few percent of the total PAR energy absorbed by the leaf [84].

### Quantitative Description of PAR in Vegetation Canopies

**Radiative Transfer in Plant Canopies** When dealing with PAR in vegetation canopies [85, 86], it is assumed that the only scatterers are plant leaves (or needles, shoots, etc.), that radiation originates from the sun only, and that thermal emission can be ignored. Thermal emission in the PAR waveband is indeed negligible (nonexistent for all practical purposes) at temperatures suitable for photosynthesis. The existence of fluorescence by green leaves, an emission source concurrent with photosynthesis, assumes the presence of incident PAR. The energy contribution of fluorescence is small compared to that of scattered PAR, and is generally masked by scattered radiation [87].

When using RTT, the optical thickness of a canopy is often described by its leaf area index (LAI): one-sided leaf area (or half of the total leaf area for plants with non-flat leaves) per unit ground area. Quite commonly,

the downward cumulative LAI,  $L(z)$ , calculated as LAI above height  $z$ , is used instead of the geometric vertical coordinate  $z$ . At the top of the canopy,  $L(z_{\text{top}})$  equals 0, then increases with depth inside the canopy. Finally, below the plant layer,  $L(0) = \text{LAI}$ .

**Radiation Interception** When RTT is applied to describe the attenuation of radiation in vegetation, a well-known result is obtained. In an environment where the scattering elements fill a volume uniformly and randomly, and are infinitesimally small, the radiance decreases exponentially:

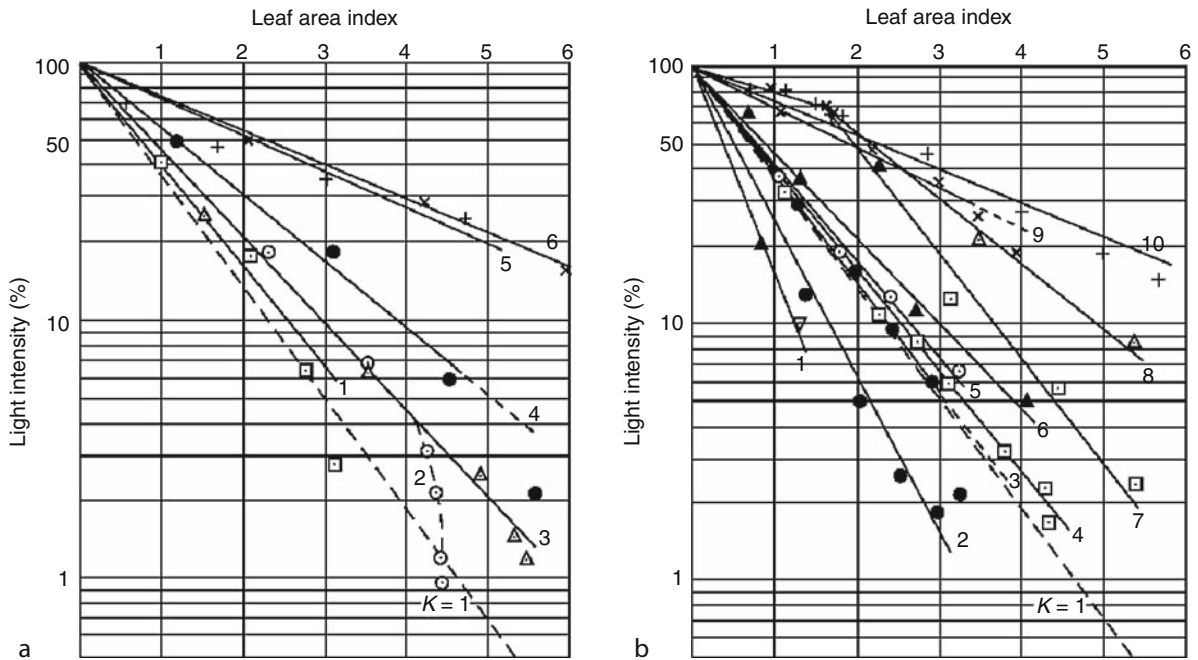
$$R_{\text{PAR}} = R_{\text{PAR}}(z_{\text{top}})e^{-kx}, \quad (8)$$

where  $R_{\text{PAR}}(z_{\text{top}})$  is the radiance before entering the scattering medium,  $k$  is the attenuation coefficient, and  $x$  is the distance from the point where radiation enters the medium (depth inside the canopy). The exponential decay described by Eq. 8 is commonly known as Beer's law, sometimes called Beer–Lambert's or Beer–Lambert–Bouguer's law. The terms attenuation and interception are used interchangeably: interception of radiation attenuates the unscattered radiation field.

The tradition of using Beer law for radiation transmission in vegetation canopies consisting of flat leaves started with the classic work by Monsi and Saeki [88, 89]. They plotted the logarithms of light transmittance of vegetation canopies during overcast days against the LAI of the canopy, and obtained straight lines (Fig. 8). They explained the variations in the slopes of the lines (i.e., attenuation coefficients) theoretically, using leaf inclination angles. Since then, Beer's law has been routinely applied to approximate PAR availability in plant canopies [90, 91].

Generally, Beer's law is exactly valid for monochromatic radiation only and is not directly applicable to radiation arriving from the whole hemisphere (i.e., irradiance  $I_{\text{PAR}}$ ). To obtain non-intercepted PAR irradiance on a horizontal surface below a plant canopy, Beer's law (Eq. 8) has to be integrated over the upper hemisphere:

$$I_{\text{PAR}} = \int_0^{\pi/2} \int_0^{2\pi} R_{\text{PAR}}(z_{\text{top}}, \vartheta, \phi) e^{-k(\vartheta, \phi)x(\vartheta, \phi)} \times \cos \vartheta \sin \vartheta \, d\theta d\phi, \quad (9)$$



**Photosynthetically Active Radiation: Measurement and Modeling. Figure 8**

Light intensity–leaf area index curves of some plant communities measured under mostly overcast conditions on different days (Reprinted from [89]). (a) The communities of the Kirigamine montane meadow; (b) the communities of the Tazima meadow and in the vicinity of Tokyo. Light intensity, a good proxy for PAR irradiance, decreases near-exponentially (See the original article for more details. Published with permission from Oxford University Press)

where  $\vartheta$  and  $\phi$  are the zenith and azimuth coordinates describing a direction in the upper hemisphere, respectively. The distance from the bottom to the top of the canopy in the direction  $(\vartheta, \phi)$  depends mostly on the zenith angle  $\vartheta$ . For a canopy layer of uniform thickness,  $x(\vartheta, \phi) = 1/\cos \vartheta$ . Equation 9 is identical to the equation relating irradiance and radiance, Eq. 4. The former can be obtained from the latter by expressing radiance as a function of canopy transmittance (which is Beer's law for vegetation canopies, Eq. 8) and considering that the differential of a solid angle can be expressed in polar coordinates as  $d\Omega = \sin \vartheta d\theta d\phi$ . Computations involving numerical integrations of canopy interception over the upper hemisphere similar to Eq. 9 are common in optical estimations of LAI [92, 93].

In most applications of RTT, the probability of a scattering event, and thus the extinction coefficient  $k$ , does not depend on the direction  $(\vartheta, \phi)$  of photon travel. However, this does not generally hold in a vegetation canopy of flat leaves. For example, in a canopy consisting of mainly horizontal leaves, the

probability of hitting a leaf is much larger for a photon traveling in the vertical direction than for one traveling in the horizontal direction. Thus, the Ross–Nilson G-function (a function returning a value between zero and one for every direction) is used to describe the effect of foliage orientation: it equals the ratio of leaf area projected in a given direction  $(\vartheta, \phi)$  to the total leaf area [85].

The distribution of foliage inside a natural canopy is not uniform and the assumptions of Beer's law are not fulfilled. To accurately describe the attenuation of radiation (including PAR) inside vegetation, more complex methods have to be used, and one must take into account the structure of the vegetation layer. In addition to the above-mentioned angular distribution of foliage, additional variables describing the geometric properties of plants at various levels (shoot, branch, crown, etc.) have to be used. Introduction of larger-scale structures decreases interception at the top of the canopy and increases it in middle canopy layers [94, 95].

**Scattering Inside the Canopy** The small fraction of photons in the PAR waveband not absorbed when hitting a leaf give rise to a phenomenon commonly known as beam enrichment. After one or two scatterings in the canopy (very few photons in the PAR waveband survive more), photons may be inserted back into the beam traveling in the direction of the receiver. In other words, as plant leaves are not black (completely absorbing), the downward flux of radiation is “enriched” by scattered photons. Naturally, the amplitude of this effect depends on the scattering properties of the scattering elements, leaves, and needles.

To describe the reflectance and transmittance properties of plant leaves, they are usually assumed to be Lambertian surfaces – the angular distribution of reflected (or transmitted) radiance does not depend on the direction of scattering. Actual leaves deviate from Lambertian surfaces, mainly due to the specular (mirror-like) reflectance from the wax coating. However, for practical purposes, there is no information that the assumption of Lambertian scattering would lead to considerable errors [96]. Therefore, the scattering properties of flat leaves are generally described by up to three numbers: the two leaf reflectance values for the abaxial and adaxial leaf sides, and one leaf transmittance (the general reciprocity relations require for the two sides of a Lambertian surface to have identical transmittance [95]). However, quite often reflectances are not available separately for the two leaf sides and the reflectances of the adaxial and abaxial sides are taken equal.

The reflectance properties of leaves depend somewhat on species, growing conditions, and leaf status. Some approximate values are given in the literature. In his seminal book, Ross used the values 0.06 and 0.09 for leaf reflectance and transmittance in the PAR region, respectively [85]. Relatively little between-species variability in leaf optical properties, compared to within-species variability, was found during an extensive study in Texas, USA [97], indicating that the leaf optical properties in this spectral region are dynamically stable along a pronounced climate gradient. For trees and shrubs, a leaf reflectance of 0.09 (standard deviation 0.01) and a leaf transmittance of 0.06 ( $\pm 0.03$ ) was proposed; for grasses the resulting numbers were 0.12 ( $\pm 0.01$ ) for reflectance and 0.06 ( $\pm 0.02$ ) for transmittance [97]. Although the measurements were performed in the spectral interval of channel 1 of the

AVHRR satellite sensor (550–700 nm) which corresponds to the green–red part of PAR, they can also be used to approximate the leaf optical properties in the whole PAR waveband with reasonably small errors.

**Radiation Field Inside a Vegetation Canopy** Canopy photosynthesis depends not only on the amount of available PAR, but also on how the irradiance is distributed: high and low PAR irradiance levels have different photosynthetic potentials. Without going into further details, it is possible to divide the locations inside a plant canopy into three groups.

1. Full sunlight. In areas inside the canopy where the sun is completely visible, or “sunflecks,” the radiation field is strongly dominated by the direct solar radiation beam. Under natural conditions, one can safely ignore the contribution of scattered PAR and assume that the spectral distribution of radiation is identical to that above the canopy.
2. Penumbra. Due to the nonzero diameter of the solar disc, shadows cast by sunrays do not have sharp edges. Full sunlight and complete shadow are always separated by a narrow strip with smoothly varying irradiance. If the angular dimensions of the shadowing object are smaller than the apparent diameter of the solar disc, for example, the object is far from the receiver, no complete shadowing can occur. Depending on the fraction of the solar disc visible, the PAR flux and spectrum in penumbra may be close to what they would be under either direct sunlight or complete shadow.
3. Shadow (umbra). Behind large objects or sufficiently deep down into the canopy, the direct solar irradiance can be considered zero. Thus, the radiation conditions in umbra are determined by the possible presence of diffuse sky radiation, radiation scattered by plant elements, and radiation reflected by the underlying soil. The spectral composition of radiation depends on skylight conditions, the amount of visible sky, and the spectral properties of canopy elements. On an overcast day, the whole canopy is effectively in a shadow cast by clouds.

The division as given above is just one of the possible approaches to categorizing the radiation field. No standard practice has emerged yet in the scientific literature: the word “sunfleck” is the general term used to

describe areas with increased irradiance, whereas the penumbra is often ignored. Depending on application, the penumbra may be treated as either an area where the sun is obstructed (i.e., shadow) or, in contrast, an area where the irradiance is above the threshold determined by the reading below a dense canopy (i.e., sunfleck).

Due to the dynamic nature of sunflecks, the radiation field inside a canopy can be highly variable, both spatially and temporally [98]. Variations in the diffuse field are generally much smaller, and thus global irradiance in umbra is much easier to measure. For this reason, canopy transmittance measurements made on a cloudy day are much more representative. For example, 412 radiometers would be required to estimate the instantaneous downward radiation flux in a pine stand with a maximum error of 10% in the midday flux above the canopy, whereas just one instrument is needed for a full-day average in a hardwood canopy [99]. When averaged over the path, the sun follows on a day or a growing season, the mean transmittance of direct radiation can be approximated by a single measurement of transmittance of diffuse sky radiation. However, due to natural limitations in the solar elevation and azimuth angles which are specific to each geographic location, systematic errors may occur [90].

Even if one has sufficient data to quantify the average PAR irradiance inside and below a vegetation canopy, this will not suffice for photosynthesis modeling purposes. It has been known for a long time that the photosynthetic response of a leaf is not linear with PAR irradiance (see section “[PAR in Vegetation Canopies](#)”). Thus, the radiation field is commonly described as composed of two fluxes, direct and diffuse PAR, and PPF values at leaf surfaces are calculated separately for the two fluxes [100]. Beside atmospheric conditions, the availability of direct solar radiation depends only on canopy structure. In contrast, diffuse radiation is generated via more numerous mechanisms.

The first source of what can be called diffuse radiation, or radiation of diminished radiance compared with direct solar radiance above the vegetation canopy, is penumbra. Penumbral effects are purely geometric and are most evident in a tall canopy of small scatterers, for example, in a needle-leaf boreal forest. At high latitudes, low sun angles further increase the pathlength of direct solar beam in the canopy, thus

making the penumbra dominate the radiation field on a clear day. In a canopy consisting of shoots, the penumbral effect alters the irradiance distribution, but also vertically redistributes the photosynthetic potential inside the canopy [101]. Coupling the nonzero angular diameter of the sun and the three-dimensional structure of the vegetation canopy can lead (at least in model calculations) to an increase in canopy photosynthetic capacity by tens of percent [102, 103]. Such computations are rare in the scientific literature since the correct prediction of penumbral irradiation assumes nonzero scatterer sizes, which is beyond the scope of traditional RTT.

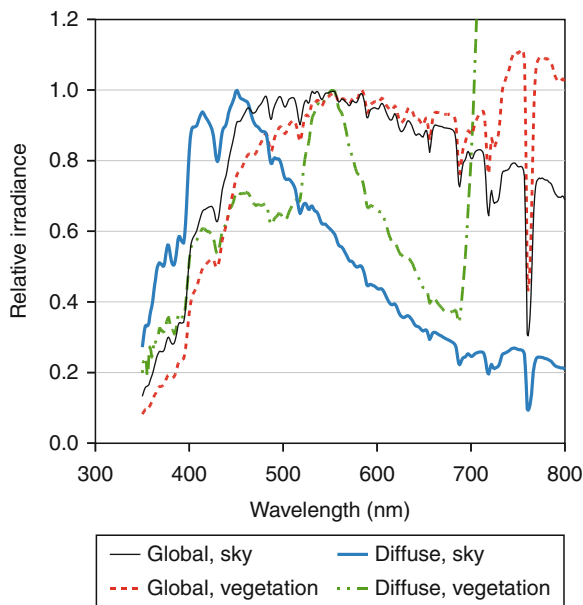
The second source of diffuse radiation, multiple scattering inside vegetation, depends on the canopy structure and cannot be easily predicted without a radiative transfer model. However, whereas modeling penumbra requires special consideration for both the spatial distribution and the dimensions of canopy elements, the dimensions of leaves are usually ignored when calculating multiple scattered radiation fluxes. Assuming infinitesimally small leaves makes it possible to apply the traditional methods for solving RTT [85, 86].

The third component of the diffuse PAR field inside a vegetation canopy is contributed by diffuse sky radiation: some of the photons scattered by the atmosphere can penetrate the vegetation layer without being intercepted by canopy elements. Depending on cloudiness and other atmospheric conditions, the diffuse sky irradiance can vary within large limits. The complications related to calculating this PAR component originate in the difficulties of correctly estimating the variability of above-canopy diffuse radiance. In contrast, the canopy transmittance can be modeled from simple structural assumptions or measured from hemispheric photography.

The diffuse sky PAR irradiance is usually less than one-third of the global PAR irradiance (section “[PAR Below the Atmosphere](#)”). Considering that not all of the sky is visible inside the canopy, the diffuse PAR irradiance (on a horizontal surface) is at least one order of magnitude smaller than the direct PAR irradiance in a sunfleck. The adaptation of leaves in deeper canopy layers (where sunflecks are rare) to low irradiances makes diffuse sky radiation more effective in inducing photosynthesis than direct PAR [104]. In other words, the light use efficiency is generally larger

for diffuse radiation than for direct irradiance: this effect is sometimes termed diffuse-radiation fertilization. Thus, an increase in diffuse PAR can lead to increased carbon assimilation. This mechanism has been proposed as the explanation of the decrease in global CO<sub>2</sub> concentration after Mt. Pinatubo's eruption in 1991. Volcanic eruptions are known to increase the amount of atmospheric aerosol for several years thus enhancing diffuse sky radiation [105, 106] and global carbon assimilation. The adverse effect of decreasing the amount of anthropogenic sulfate aerosol could similarly lead to a decrease or fall in global photosynthesis [107].

The spectral distribution of PAR above and below a vegetation canopy is shown in Fig. 9. Four spectral distributions are shown, corresponding to global and diffuse radiation above and below a closed alder canopy (LAI = 2). The distributions are normalized so that their maximum value in the PAR waveband equals unity. Whereas changes in the spectral distribution of global PAR are relatively small, alteration of the spectral



**Photosynthetically Active Radiation: Measurement and Modeling. Figure 9**

Spectral distribution of radiation on a clear day above and below a vegetation canopy. Measurements were made inside and next to a gray alder (*Alnus incana*) plantation (leaf area index  $L = 2.1$ ) in Tõravere, Estonia

distribution for the diffuse field is quite significant. Above the canopy, the diffuse radiation spectrum peaks at blue wavelengths. The diffuse PAR field below a canopy reaches its maximum at about 550 nm, which corresponds to green light. Further, in the near-infrared (NIR) region, at wavelengths immediately above 700 nm, the diffuse spectral irradiance below the canopy increases to levels much higher than those in the PAR waveband. Such an increase emphasizes the requirement to consider spectral errors with care: if the sensor sensitivity cutoff at 700 nm is not sharp enough, measurements of diffuse PAR inside a vegetation canopy are contaminated by the high NIR irradiance. Also, the differing spectral composition of diffuse radiation between the higher and lower canopy layers must be taken into account when calculating the contribution of diffuse radiation to canopy photosynthesis.

Explicit treatment of the total PAR available for photosynthesis is thus a complex task [98]. Exact computations require that many factors are taken into account: dimensions of the solar disc and the scattering elements, detailed structure of the vegetation canopy, spectral properties of leaves, nongreen canopy elements and soil, spectral and angular distributions of incident radiation, etc. Due to the large spatial and temporal variability of PAR inside a plant canopy, its measurement and empirical analysis are also extremely complicated and only a few examples are presented in the literature [20, 108].

### Measurement of PAR Absorbed by Canopies

**Direct Measurement of APAR** The PAR flux absorbed by plant canopies (APAR) may be either measured directly using PAR sensors or estimated indirectly, based on canopy gap fraction measurements. Leaf area index (LAI) measurements may also be used to estimate the PAR absorbed by the canopy. First, a description of direct PAR measurements is given.

The PAR absorbed by vegetation equals the total amount of radiative energy absorbed by all plant surfaces and can thus be measured in either quantum or energy units. The formulation of the problem is identical for the two representations. Here, the quantum APAR ( $Q_{PAR}^4$ ) is used as an example. Due to the different spectral compositions of PAR that is incident or reflected and transmitted by a vegetation canopy,

converting between  $Q_{PAR}^A$  and its companion quantity in energy units,  $I_{PAR}^A$ , is not a straightforward task. Thus, estimates of fAPAR from quantum measurements are not directly comparable with fAPAR values calculated in energy units. However, considering the measurement uncertainties and the errors related to retrieval of fAPAR from remote sensing measurements, the differences are ignored here.

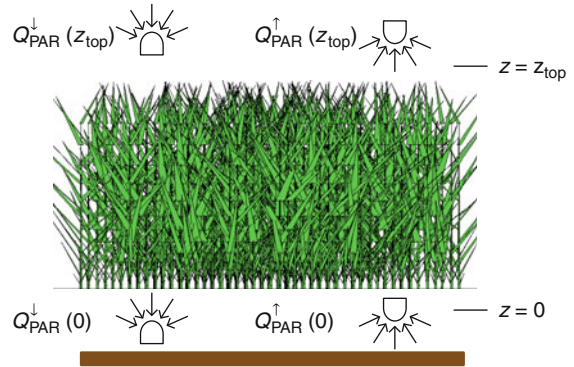
$Q_{PAR}^A$  is computed from the energy conservation law in the canopy using hemispherical fluxes:

$$Q_{PAR}^A = Q_{PAR}^{\downarrow}(z_{top}) - Q_{PAR}^{\uparrow}(z_{top}) - Q_{PAR}^{\downarrow}(0) + Q_{PAR}^{\uparrow}(0). \quad (10)$$

where  $Q_{PAR}^*(z)$  represents downward  $Q_{PAR}^{\downarrow}(z)$  or upward  $Q_{PAR}^{\uparrow}(z)$  hemispherical PAR fluxes at the bottom ( $z = 0$ ) or top ( $z = z_{top}$ ) of the canopy. Therefore, the hemispherical PAR sensors should be precisely located to obtain representative measurements of the several terms in the PAR balance:

- $Q_{PAR}^{\downarrow}(z_{top})$ , the incident PAR at the top of the canopy is measured using upward-looking sensors at the top of the canopy
- $Q_{PAR}^{\uparrow}(z_{top})$ , the reflected PAR is measured using downward-looking sensors at the top of the canopy
- $Q_{PAR}^{\downarrow}(0)$ , the transmitted PAR is measured by upward-looking sensors placed at the bottom of the canopy
- $Q_{PAR}^{\uparrow}(0)$ , the PAR reflected by the soil is measured using downward-looking sensors at the bottom of the canopy

The number of sensors used to measure the several terms in a representative way depends mainly on the heterogeneity of the canopy and on the typical footprint of the sensor. About half of the flux collected by a hemispherical sensor comes from inside a circle whose radius equals the distance of the sensor from the object it is looking at [109]; about 96% of the signal originates from inside a circle with a radius of five times the distance. Therefore, many sensors (between 5 and 50) are required to properly measure all the terms in Eq. 10 at the bottom of the canopy: the distance from the sensor to the bottom of the canopy (when measuring  $Q_{PAR}^{\downarrow}(0)$ ) or the soil (when measuring  $Q_{PAR}^{\uparrow}(0)$ ) is generally limited. Conversely, just one sensor is



### Photosynthetically Active Radiation: Measurement and Modeling. Figure 10

Configuration of PAR sensors to measure APAR and fAPAR. Sensor A measures the incident PAR ( $Q_{PAR}^{\downarrow}(z_{top})$ ), sensor B measures the reflected PAR ( $Q_{PAR}^{\uparrow}(z_{top})$ ), sensor C measures the transmitted PAR ( $Q_{PAR}^{\downarrow}(0)$ ), and sensor D (usually omitted) measures the soil reflectance ( $Q_{PAR}^{\uparrow}(0)$ )

required for the incoming PAR,  $Q_{PAR}^{\downarrow}(z_{top})$ , and only a few for the reflected PAR,  $Q_{PAR}^{\uparrow}(z_{top})$ , depending on the distance between the sensor and the top of the canopy (Fig. 10).

The fraction of absorbed PAR (fAPAR,  $F_{PAR}^A$ ), may be derived from Eq. 10 by dividing all the terms by the incident PAR:

$$F_{PAR}^A = 1 - r_{CAN} - t_{CAN}(1 - r_{SOIL}), \quad (11)$$

where  $r_{CAN}$  is the reflectance of the canopy,  $t_{CAN}$  is the transmittance of the canopy, and  $r_{SOIL}$  is the soil reflectance. As mentioned above,  $F_{PAR}^A$  is used here to denote the quantum fAPAR, or the absorbed fraction of incident photons in the PAR waveband. Note that all these variables are bihemispherical quantities [110, 111], although the directional integration of incident radiation requires, as a minimum, weighing the direct and diffuse components of canopy transmittance. According to Eq. 11, fAPAR is a very convenient quantity: it is independent on the magnitude of the incident irradiance. However, fAPAR is somewhat sensitive to the irradiance geometry: the attenuation of the direct component of the incident radiation field varies strongly with the direction of the sun as well as canopy characteristics.



If the soil reflectance  $r_{SOIL}$  is assumed to be known, the PAR balance may be approximated by using measurement data for the three first terms only, that is, the incident ( $Q_{PAR}^{\downarrow}(z_{top})$ ), reflected ( $Q_{PAR}^{\uparrow}(z_{top})$ ), and transmitted ( $Q_{PAR}^{\downarrow}(0)$ ) radiation. Equation 11 may further be simplified as in the PAR domain, where very little multiple scattering is expected,

$$F_{PAR}^A \simeq (1 - t_{CAN})(1 - r_{\infty}), \quad (12)$$

where  $r_{\infty}$  is the asymptotic value of canopy reflectance when the leaf area index,  $L$ , tends toward infinity. Values of  $r_{\infty}$  are generally small in the PAR domain (around 0.06 [112]) since most photons are absorbed by the green leaves. In these conditions, the PAR balance, Eq. 10, may be approximated by measuring only the incident ( $Q_{PAR}^{\downarrow}(z_{top})$ ) and the transmitted ( $Q_{PAR}^{\downarrow}(0)$ ) terms. It is thus possible to avoid problems in measuring the soil reflectance  $r_{SOIL}$ : as highlighted earlier, many sensors are required due to the small distance between soil and sensor and the possible effect of the sensor shadow in its footprint. It also avoids problems related to the dependence of  $r_{SOIL}$  on the directionality of incident radiation.

Because the PAR balance and thus fAPAR both depend on the varying illumination geometry, continuous measurements are required. Furthermore, fAPAR measurements are generally used in light-use efficiency models [3, 113] which require daily and even seasonally integrated fAPAR values. The PAR radiance measurement system thus needs to be set in place from several days up to several months. In practice, this calls for weatherproof systems with sufficient autonomy both in terms of energy and memory. Affordable systems meeting these requirements and able to replicate individual observations for improved spatial sampling have been developed only recently. However, instantaneous measurements using several view directions may be achieved by different existing systems, allowing the reconstruction of fAPAR values for any (possibly modeled) illumination geometry.

In all situations, the approach based on radiation balance assesses the value of PAR absorbed by the canopy, independently of the nature of the radiation intercepting elements. As a consequence, when

non-photosynthetic material (such as trunks, branches, or senescent leaves) constitutes a significant fraction of canopy, the true fAPAR, or PAR absorbed by the green photosynthetically active elements, is overestimated.

**Estimation of fAPAR from fIPAR** Green leaves generally absorb a very large fraction of light in the PAR domain, that is, they appear almost black from a pure radiative standpoint. Considering this, the fraction of absorbed PAR may be approximated by the fraction of intercepted PAR (fIPAR,  $F_{PAR}^I$ ):

$$F_{PAR}^A = 1 - t_{CAN}. \quad (13)$$

Combining Eqs. 12 and 13 provides a relation between the fraction of absorbed PAR and the intercepted fraction:

$$F_{PAR}^A = F_{PAR}^I(1 - r_{\infty}).$$

The validity of this approximation has been extensively investigated and found to hold with reasonable accuracy [114–117].

## Instruments for Measuring fAPAR

**Directional and Flux Measurements** Based on the directionality of measurements, fAPAR sensors can be divided into two subgroups. The first group contains instruments that disregard the directional distribution of incident PAR (either by integrating over the upper hemisphere or looking into only one particular direction); the second group consists of instruments measuring with a field of view divided between different directions (multidirectional devices). Generally, the lack of directional sampling by instruments in the first group must be compensated by increased spatial sampling.

Ceptometers are devices consisting of an array of hemispherical sensors aligned on a single support, allowing for spatial representativeness. They are particularly well suited for crops: a ceptometer covering a transect representative of a forest canopy would be too long to be moved easily between the measurement locations. While a ceptometer is used to measure radiation below the canopy, the incident PAR can be simultaneously recorded with an additional PAR sensor. Examples of such sensor arrays are AccuPAR (Decagon, USA), SunScan (Delta-T, UK), and PAR/LE (Solems, France).

Other specialized multipoint radiation measurement systems have been developed (for mainly in-house use) on various scales. For example, spatial distributions of the radiation field (with focus on PAR) affecting a conifer shoot have been measured using optical fibers and a CCD matrix [118]. At the other end of the scale are systems capable of characterizing tree-level heterogeneity using hundreds of sensors attached to 5-m-long booms [119]. PAR @METER [120] is a recently developed device capable of continuously monitoring the transmitted PAR at different points inside and above a vegetation canopy. Incident and transmitted PAR are simultaneously recorded and stored in a network of sensors placed according to a predefined spatial sampling scheme using wireless computer connections.

Directional devices are generally less common than the hemispherical instruments listed above. Examples of directional, below-canopy radiation measuring devices are TRAC (Natural Resources, Canada) and DEMON (CSIRO, Australia). They measure direct sunlight and use different approximations to characterize the plant architecture. TRAC inverts the light transmittance profiles obtained on a transect based on a model of canopy gap size distribution [121]. It accounts for the nonhomogeneous distribution of foliage in certain canopies (also called clumping effect [122]) by inverting the measured sunfleck length distribution. DEMON makes use of Beer's law and a special zenith angle at which canopy transmittance does not depend on leaf orientation [123] to retrieve LAI from incident and transmitted PAR measurements.

The instruments mentioned above are designed for measurements of transmitted PAR. However, by adding canopy reflectance measurements, the complete PAR balance can be obtained (Fig. 10). From such balance measurements, it is also possible to calculate the fraction of absorbed PAR, fAPAR.

**Multidirectional Transmission Instruments** The incoming PAR may be decomposed into the direct component coming from the sun and the diffuse component due to light scattering in the atmosphere. For each of those components, a fAPAR value can be associated. The total fAPAR can then be written as:

$$F_{PAR}^A = (1 - f_{diff}) F_{PAR}^A(\Omega_S) + f_{diff} F_{PAR}^A(2\pi^+), \quad (14)$$

where  $f_{diff}$  is the diffuse fraction of radiation in global irradiance,  $\Omega_S$  is the direction of the sun, and  $2\pi^+$  is used to denote the upper hemisphere;  $F_{PAR}^A(\Omega_S)$  and  $F_{PAR}^A(2\pi^+)$  are thus the fAPAR values for direct-only or diffuse-only incidence, respectively. Sometimes,  $F_{PAR}^A(\Omega_S)$  is called the black-sky fAPAR and correspondingly,  $F_{PAR}^A(2\pi^+)$  the white-sky fAPAR. To apply Eq. 14 to the calculation of  $F_{PAR}^A$  for all sky conditions, the directional characteristics of  $t_{CAN}(\Omega)$  have to be known.

Directional devices provide measurements of canopy transmittance,  $t_{CAN}(\Omega)$ , in a number of directions  $\Omega = (\vartheta, \phi)$ . Two types of devices are mainly used: the LAI-2000 instrument [124] and digital hemispherical cameras (DHC) [125, 126]. Lidar systems may also access the directional variation of light transmittance, although the technique might be better suited for other applications related to detailed characterization of canopy architecture. The LAI-2000 instrument measures light transmitted in the blue wavelengths to the bottom of the canopy in five concentric rings of  $15^\circ$  in the range  $0 < \vartheta < 70^\circ$ . For each ring, all azimuths directions are accounted for. Measurements are generally taken under diffuse conditions to prevent an unwanted sensitivity to the specific sun direction, while minimizing any possible sun glint on the leaves. The blue spectral region is used since, at these wavelengths, leaves appear almost black and diffuse sky scattering is at its peak. The view azimuth angle can be modified using a series of view-limiting caps to block out a part of the sky or focus the measurements toward specific directions of interest.

Hemispherical cameras provide estimates of the gap fraction over the whole hemisphere. If the angular distribution of incident radiation is known, the gap fraction may be converted into canopy non-interceptance. Again, assuming that leaves are black at the visible wavelengths used by cameras, the canopy interceptance is converted into canopy absorptance. DHC usually involves a high-resolution digital camera and an attached fisheye lens. This fisheye lens projects the whole upper hemisphere onto the digital array of the camera, producing circular images. However, historical data recorded on black-and-white film may still be encountered and, in some cases, photographs made using ordinary lens are used (i.e., any lens not covering the whole hemisphere). The a posteriori processing of digital images provides the fraction of the upper

hemisphere covered by vegetation. This is done by classifying each pixel as sky or non-sky, that is, applying a threshold to divide the pixels constituting the image into two classes [126].

Furthermore, the variations in vegetation canopy transmittance with zenith (and sometimes also azimuth) angle may be used to reconstruct the diurnal variation of fAPAR. Note that, similarly to sensors measuring the transmitted PAR, no distinction is made between green photosynthetically active elements and the non-photosynthetic material. This may lead to an overestimation of the actual value of the true fAPAR. However, when using hemispherical photographs taken from above canopies, it may be possible to distinguish between green and nongreen elements. Unfortunately, downward looking photography is limited to relatively short canopies for obvious practical reasons. DHC techniques are very efficient by allowing instantaneous measurements that can be replicated multiple times to improve the spatial sampling while accessing the diurnal variation of fAPAR. However, such measurements are only representative of the current canopy architecture, and measurements should be repeated along the growing season to match the canopy architecture dynamics.

A novel approach consists in using digital cameras as hemispherical radiation receivers [127, 128]. Similarly to the traditional DHC approach, a fish-eye lens projects the whole hemisphere onto the sensor array. However, instead of just applying a threshold to identify the gaps in the canopy, the new “calibrated camera” method treats the receiving surface of the camera as a two-dimensional array of miniature quantum receivers. Each array element receives radiation from a single direction in the upper hemisphere. The spectral sensitivity functions of the array elements have maxima in the optical region of electromagnetic radiation, that is, in the PAR waveband. Therefore, after proper laboratory calibration, a raw digital image stored in the camera can be treated as a (PAR) radiance measurement result. However, these measurements must be treated with care because modern consumer cameras are complex optical systems designed for producing visually good-looking images, not recording spectral radiance values.

**Relationships Between fAPAR and LAI** The devices described above are often used (or even designed) for

measuring canopy leaf area index (LAI) [92, 93]. Generally, all techniques to estimate LAI from PAR transmittance measurements rely on Beer’s law (Eq. 8). The directional instruments allow for a more accurate integration over the hemisphere required for the accurate application of Beer’s law. However, a direct use of it, without spectral integration, is still quite common when relating PAR irradiance and LAI [129]. The opposite link is also often made: if the canopy LAI is known, fAPAR can be modeled using the known optical properties of the elements constituting the vegetation canopy and some basic knowledge of canopy structure. All these calculations are based on the Beer’s law specially formulated for vegetation canopies, as described in section “Quantitative Description of PAR in Vegetation Canopies”.

### APAR and fAPAR from Satellite Observations

The fraction of absorbed photosynthetically active radiation, fAPAR was probably the first biophysical variable to be estimated from remote sensing observations from NDVI, the normalized difference vegetation index computed as  $NDVI = \frac{r_{NIR} - r_{RED}}{r_{NIR} + r_{RED}}$ , where  $r_{NIR}$  and  $r_{RED}$  are the top-of-canopy reflectance in the near-infrared (NIR) and red (RED) bands, respectively [130]. The early empirical relationships were later explained by investigating the radiative transfer in canopies [112, 131]. Compared to other biophysical variables (such as LAI), fAPAR appears to be retrievable much more accurately and robustly [132, 133]. The optimal configuration for retrieving fAPAR includes four spectral bands: red, near-infrared, green, and red edge. Simple observations in the red and near-infrared in view directions close to nadir were found to lead to slightly degraded performance.

The optimal view angle for a satellite instrument is not directly down (nadir), but in the principal solar plane close to the hot spot (the direction of the sun, corresponding to backward scattering), and in the perpendicular solar plane at zenith angles close that of the sun [132, 133]. Alternatively, directions around 60° from nadir in the backscattering direction or in the perpendicular plane were also shown to be close to optimal [134]. However, as these optimal configurations are not generally available, most algorithms for fAPAR retrieval focus on the minimization of

directional effects by simply making use of whatever remote sensing data can be obtained.

In addition to NDVI, indices have been developed to correct for the contribution of soil to the measured reflectance, or to use the more readily available top-of-atmosphere reflectance instead of the top-of-canopy value [116, 134]. Further, look-up tables have been used to derive fAPAR from MODIS top-of-canopy reflectance observations after calibration with radiative transfer model simulations [135]. However, when the physically based algorithm (known as the main

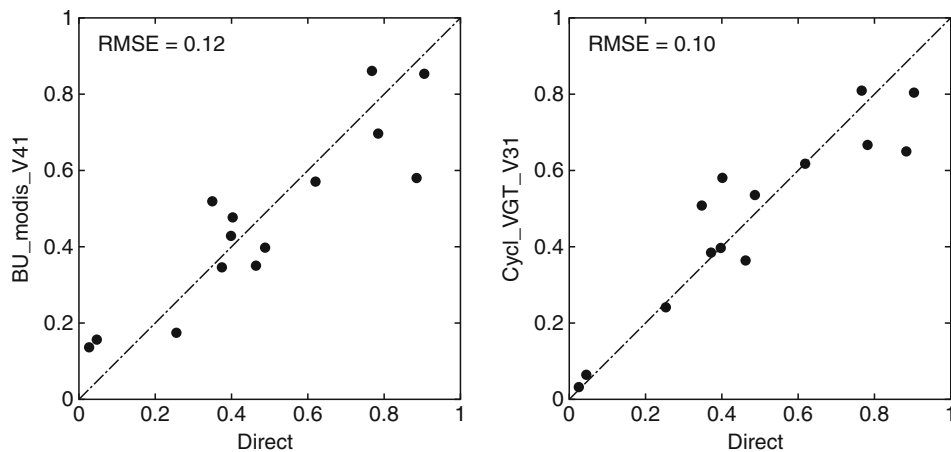
algorithm) fails, a backup algorithm is triggered using relationships between fAPAR and MODIS NDVI. Neural networks have also been used [120, 136] to operationally retrieve fAPAR from satellite-measured radiances.

The main fAPAR products derived from satellite observations (Table 2) thus demonstrate a wide range of either empirical or physically based approaches. To obtain these fAPAR products, the needed inputs are either top-of-canopy or top-of-atmosphere reflectance values observed in 2–13 reflectance bands. Some

**Photosynthetically Active Radiation: Measurement and Modeling. Table 2** Examples of fAPAR values derived from satellite-based reflectance measurements

Product name	Approach	Sensor	Reference
NDVI	Empirical linear regression	AVHRR	[130, 137, 140]
NDVI	Linear regression of RT simulations	AVHRR	[112, 141]
RDVI	Linear regression of RT simulations	POLDER	[134]
JRC-FAPAR	VI calibrated using RT simulations	PARASOL, SEVIRI	[116]
TOC-VEG	NN calibrated using RT simulations	MERIS	[136]
TOA-VEG	NN calibrated using RT simulations	MERIS	[142]
MODIS	LUT from RT simulations	MODIS	[135]
CYCLOPES	NN calibrated using RT simulations	Vegetation	[120]
GLOBCARBON	Derived from LAI product	Vegetation, MERIS, AATSR	[143]
GEOLAND2	NN calibrated using other products	Vegetation	[144]

VI vegetation index, RT radiative transfer, NN neural network, LUT look-up table



**Photosynthetically Active Radiation: Measurement and Modeling. Figure 11**

Comparison of ground-measured (horizontal axis) and satellite (vertical axis) estimates (MODIS: left, CYCLOPES: right) of fAPAR

algorithms use a priori information on vegetation type or rely on a land cover map. Most of the algorithms provide an instantaneous black-sky fAPAR value at the time of satellite overpass, while a few others use multidirectional observations, to provide a daily integrated black-sky value. Note that most of the polar orbiting sensors considered here are on satellites in sun-synchronous orbits with equatorial crossing time close to 10:00 local time. Under these conditions, the instantaneous black-sky fAPAR value is a good approximation of the daily integrated fAPAR value.

Individual validation exercises have been reported by several authors [116, 137–139]. They generally show a reasonable agreement between ground-measured fAPAR and satellite estimates, with RMSE values around 0.10–0.15 (in fAPAR units) (Fig. 11). Considering the complex interactions between radiation and vegetation canopies described above, fAPAR also has a most desirable feature: it is almost independent of scale. Values of fAPAR derived from algorithms applied at higher spatial resolution and integrated over a coarser spatial domain provide similar values to those derived using the same algorithm applied directly to the coarser spatial resolution [133]. Unfortunately, the same cannot be said of any other vegetation parameter derived from remote sensing data.

### Future Directions

The current research related to PAR measurement (and, inevitably, modeling) is aimed at utilizing the technological advances in (remote) sensing technology to better characterize the environment we live in. Photosynthesis is the energy source for all life on earth. The raw energy for life is originally dispersed in the form of electromagnetic radiation arriving from our closest star. Although the importance of photosynthesis, and the role of shortwave radiation in it, has always been acknowledged, there are still large gaps in our understanding.

From a more technical point of view, the most evident and surprising gap is a lack of comprehensive ground-based measurement network. Fortunately, this lack of basic monitoring does not result in severe ignorance of global PAR availability. This is evidenced by the ongoing satellite measurements and the

simultaneous model developments – to convert satellite sensor readings into radiation fluxes absorbed by vegetation hundreds of kilometers below. The progress is also witnessed by the large number of scientific articles with keywords such as fAPAR, satellite remote sensing, and global productivity. The ultimate goal of this research, however, is not only to give a detailed quantitative measure of the health of our planet, but also to provide the physical basis for describing and understanding the very fundamental links between the physical and biological environments.

### Bibliography

1. McCree KJ (1972) The action spectrum, absorptance and quantum yield of photosynthesis in crop plants. *Agric Meteorol* 9(3–4):191–216
2. McCree KJ (1972) Test of current definitions of photosynthetically active radiation against leaf photosynthesis data. *Agric Meteorol* 10(6):443–453
3. Monteith JL (1977) Climate and efficiency of crop production in Britain. *Philos Trans R Soc Lond B Biol Sci* 281(980):277–294
4. Shibbles R (1976) Terminology pertaining to photosynthesis. *Crop Sci* 16(3):437–439
5. CIE (1993) Terminology for photosynthetically active radiation for plants. CIE Collect Photobiol Photochem 106(6):42–46, ISBN 3900734461
6. McCree KJ (1965) Light measurements in plant growth investigations. *Nature* 206(4983):527–528
7. Inada K (1976) Action spectra for photosynthesis in higher-plants. *Plant Cell Physiol* 17(2):355–365
8. Barnes C, Tibbitts T, Sager J, Deitzer G, Bubenheim D, Koerner G, Bugbee B (1993) Accuracy of quantum sensors measuring yield photon flux and photosynthetic photon flux. *Hortscience* 28(12):1197–1200
9. Bonhomme R (2000) Beware of comparing RUE values calculated from PAR vs solar radiation or absorbed vs intercepted radiation. *Field Crops Res* 68(3):247–252, cited By (since 1996) 16
10. Sinclair TR, Muchow RC (1999) Radiation use efficiency. *Adv Agron* 65:215–265
11. Ross J, Sulev M (2000) Sources of errors in measurements of PAR. *Agric For Meteorol* 100(2–3):103–125
12. Gaastra P (1959) Photosynthesis of crop plants as influenced by light, carbon dioxide, temperature, and stomatal diffusion resistance. *Mededel Landbouwhogeschool Wageningen* 59:1–68
13. Nichiporovich A (1960) Conference on measurement of visible radiation in plant physiology. *Sov Plant Physiol* 7:744–747
14. McCree K (1966) A solarimeter for measuring photosynthetically active radiation. *Agric Meteorol* 3(5–6):353–366, cited By (since 1996) 20
15. Ross J (1975) Radiative transfer in plant communities. In: Monteith JL (ed) *Vegetation and the atmosphere*, vol 1. Academic, London/New York, pp 13–55

16. ISO (2004) Spatial distribution of daylight. CIE standard general sky. ISO 15469:2004/CIE 011:2003. ISO, 2004
17. LI-COR (1986) LI-COR radiation sensors. instruction manual. Publ. no 8609-56. Lincoln, Nebraska
18. LI-COR (1991) LI-COR radiation measurement instruments. Lincoln, Nebraska
19. Pearcy R (1989) Radiation and light measurements. In: Pearcy R, Ehleringer J, Mooney H, Rundel P (eds) Plant physiological ecology. Chapman & Hall, New York, pp 97–116, ch. 6
20. Ross J, Sulev M, Saarelaid P (1998) Statistical treatment of the PAR variability and its application to willow coppice. *Agric For Meteorol* 91(1–2):1–21
21. Norman JM, Tanner CB, Thurtell GW (1969) Photosynthetic light sensor for measurements in plant canopies. *Agron J* 61(6):840–843
22. Britton CM, Dodd JD (1976) Relationships of photosynthetically active radiation and shortwave irradiance. *Agric Meteorol* 17(1):1–7
23. Howell TA, Meek DW, Hatfield JL (1983) Relationship of photosynthetically active radiation to shortwave radiation in the San-Joaquin Valley. *Agric Meteorol* 28(2):157–175
24. Alados I, FoyoMoreno I, Alados-Arboledas L (1996) Photosynthetically active radiation: Measurements and modelling. *Agric For Meteorol* 78(1–2):121–131
25. Fielder P, Comeau P (2000) Construction and testing of an inexpensive PAR sensor. Ministry of Forests Research, British Columbia, Working Paper 53/2000
26. Rodskjer N, Kornher A (1971) Über die bestimmung der strahlungsenergie im wellen-längenbereich von 0, 3–0, 7 [μ] in pflanzenbeständen. *Agric Meteorol* 8:139–150
27. Slomka J, Slomka K (1986) Participation of photosynthetically active radiation in global radiation. In: Publications of the Institute of Geophysics, Polish Academy of Sciences, p 197
28. Stanhill G, Fuchs M (1977) Relative flux-density of photosynthetically active radiation. *J Appl Ecol* 14(1):317–322
29. Stigter CJ, Musabilha VMM (1982) The conservative ratio of photosynthetically active to total radiation in the tropics. *J Appl Ecol* 19(3):853–858
30. Blackburn WJ, Proctor JTA (1983) Estimating photosynthetically active radiation from measured solar irradiance. *Sol Energy* 31(2):233–234
31. Hansen V (1984) Spectral distribution of solar-radiation on clear days - a comparison between measurements and model estimates. *J Climate Appl Meteorol* 23(5):772–780
32. Rao CRN (1984) Photosynthetically active components of global solar-radiation - measurements and model computations. *Arch Meteorol Geophys Bioclimatol B Theor Appl Climatol* 34(4):353–364
33. Spitters CJT, Toussaint HAJM, Goudriaan J (1986) Separating the diffuse and direct component of global radiation and its implications for modeling canopy photosynthesis. 1. components of incoming radiation. *Agric For Meteorol* 38(1–3):217–229
34. Tooming H, Niilisk H (1967) Transition coefficients from integrated radiation to photosynthetic active radiation (PAR) under field conditions. In: Phytoactinometrical investigations of plant canopy (in Russian). Valgus Publishers, Tallinn, pp 140–149
35. Karalis JD (1989) Characteristics of direct photosynthetically active radiation. *Agric For Meteorol* 48(3–4):225–234
36. Jacovides CP, Kallos GB, Steven MD (1993) Spectral band resolution of solar-radiation in Athens, Greece. *Int J Climatol* 13(6):689–697
37. Stephens K, Strickland JDH (1962) Use of a thermopile radiometer for measuring the attenuation of photosynthetically active radiation in the sea. *Limnol Oceanogr* 7(4):485–487
38. Mims FM (2003) A 5-year study of a new kind of photosynthetically active radiation sensor. *Photochem Photobiol* 77(1):30–33
39. Möttus M, Ross J, Sulev M (2001) Experimental study of ratio of PAR to direct integral solar radiation under cloudless conditions. *Agric For Meteorol* 109(3):161–170
40. Grant RH, Heisler GM, Gao W (1996) Photosynthetically-active radiation: Sky radiance distributions under clear and overcast conditions. *Agric For Meteorol* 82(1–4):267–292
41. Grant R, Heisler G (1997) Obscured overcast sky radiance distributions for ultraviolet and photosynthetically active radiation. *J Appl Meteorol* 36(10):1336–1345
42. McArthur LJB (2004) Baseline Surface Radiation Network (BSRN) operations manual, version 2.1. Technical Report WMO/TD-No. 879, World Climate Research Programme Baseline Surface Radiation Network. [http://www.bsrn.awi.de/fileadmin/user\\_upload/Home/Publications/McArthur.pdf](http://www.bsrn.awi.de/fileadmin/user_upload/Home/Publications/McArthur.pdf)
43. Michalsky JJ, Harrison LC, Berkheiser WE (1995) Cosine response characteristics of some radiometric and photometric sensors. *Sol Energy* 54(6):397–402
44. Su WY, Charlock TP, Rose FG, Rutan D (2007) Photosynthetically active radiation from Clouds and the Earth's Radiant Energy System (CERES) products. *J Geophys Res Biogeosci* 112(G2):G02022
45. BSRN (2005) UV and PAR measurement. Report of the eighth session of the Baseline Surface Radiation Network (BSRN) workshop and scientific review meeting (Exeter, UK, 26–30 July 2004), vol 4/2005, pp 14–16. World Climate Research Programme, Informal Report 4/2005, 2005
46. Gueymard CA (2008) REST2: High-performance solar radiation model for cloudless-sky irradiance, illuminance, and photosynthetically active radiation - validation with a benchmark dataset. *Sol Energy* 82(3):272–285
47. Alados I, Alados-Arboledas L (Jan 1999) Direct and diffuse photosynthetically active radiation: measurements and modelling. *Agric For Meteorol* 93:27–38
48. Bosch JL, Lopez G, Batlles FJ (Jan 2009) Global and direct photosynthetically active radiation parameterizations for clear-sky conditions. *Agric For Meteorol* 149:146–158
49. Perez R, Ineichen P, Seals R, Michalsky J, Stewart R (1990) Modeling daylight availability and irradiance components from direct and global irradiance. *Sol Energy* 44(5):271–289
50. Alados I, Olmo FJ, Foyo-Moreno I, Alados-Arboledas L (Apr 2000) Estimation of photosynthetically active radiation under cloudy conditions. *Agric For Meteorol* 102:39–50

51. Alados-Arboledas L, Olmo FJ, Alados I, Perez M (2000) Parametric models to estimate photosynthetically active radiation in Spain. *Agric For Meteorol* 101(2–3):187–201
52. Jacovides CP, Timbrios F, Asimakopoulos DN, Steven MD (1997) Urban aerosol and clear skies spectra for global and diffuse photosynthetically active radiation. *Agric For Meteorol* 87(2–3):91–104
53. Papaioannou G, Papanikolaou N, Retalis D (1993) Relationships of photosynthetically active radiation and shortwave irradiance. *Theor Appl Climatol* 48(1):23–27
54. Jacovides CP, Tymvios FS, Asimakopoulos DN, Theofilou KM, Pashiardes S (2003) Global photosynthetically active radiation and its relationship with global solar radiation in the Eastern Mediterranean basin. *Theor Appl Climatol* 74(3–4):227–233
55. Wang Q, Kakubari Y, Kubota M, Tenhunen J (Jan 2007) Variation on PAR to global solar radiation ratio along altitude gradient in Naeba Mountain. *Theor Appl Climatol* 87:239–253
56. McCree K (1981) Photosynthetically active radiation. In: Pirson A, Zimmermann M (eds) *Encyclopedia of plant physiology*, vol 12A. Springer, Berlin, Heidelberg, pp 41–55
57. Dye DG (2004) Spectral composition and quanta-to-energy ratio of diffuse photosynthetically active radiation under diverse cloud conditions. *J Geophys Res Atmos* 109(D10):D10203
58. Kirk JTO (1994) *Light and photosynthesis in aquatic ecosystems*, 2nd edn. Cambridge University Press, Cambridge/England
59. Buiteveld H, Hakvoort JH, Donze M (Oct 1994) Optical properties of pure water. In: Jaffe JS (ed) *Society of Photo-Optical Instrumentation Engineers (SPIE) conference series*, vol 2258 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pp 174–183
60. Bricaud A, Babin M, Morel A, Claustre H (1995) Variability in the chlorophyll-specific absorption-coefficients of natural phytoplankton - analysis and parameterization. *J Geophys Res Oceans* 100(C7):13321–13332
61. Bricaud A, Stramski D (1990) Spectral absorption-coefficients of living phytoplankton and nonalgal biogenous matter - a comparison between the Peru upwelling area and the Sargasso sea. *Limnol Oceanogr* 35(3):562–582
62. Jerlov NG (1976) *Marine optics*. Elsevier, Amsterdam/New York
63. Austin RW, Petzold TJ (1986) Spectral dependence of the diffuse attenuation coefficient of light in ocean waters. *Opt Eng* 25(3):471–479
64. Reinart A, Herlevi A (1999) Diffuse attenuation coefficient in some Estonian and Finnish lakes. *Proc Estonian Acad Sci Biol Ecol* 48(4):267–283
65. Reinart A, Arst H, Blanco-Sequeiros A, Herlevi A (1998) Relation between underwater irradiance and quantum irradiance in dependence on water transparency at different depths in the water bodies. *J Geophys Res Oceans* 103(C4):7749–7752
66. Dera J (1992) *Marine physics*. Elsevier, Amsterdam
67. Ehn J, Granskog MA, Reinart A, Erm A (2004) Optical properties of melting land-fast sea ice and underlying seawater in Santala Bay, Gulf of Finland. *J Geophys Res Oceans* 109(C9):C09003
68. Morel A, Smith RC (1974) Relation between total quanta and total energy for aquatic photosynthesis. *Limnol Oceanogr* 19(4):591–600
69. Aas E (1971) Natural history of Hardangerfjord. 9. Irradiance in Hardangerfjorden 1967. *Sarsia* 46:59–78
70. Ohmura A, Dutton EG, Forgan B, Frohlich C, Gilgen H, Hegner H, Heimo A, Konig-Langlo G, McArthur B, Muller G, Philipona R, Pinker R, Whitlock CH, Dehne K, Wild M (1998) Baseline surface radiation network (BSRN/WCRP): New precision radiometry for climate research. *Bull Am Meteorol Soc* 79(10):2115–2136
71. Baldocchi D, Falge E, Gu LH, Olson R, Hollinger D, Running S, Anthoni P, Bernhofer C, Davis K, Evans R, Fuentes J, Goldstein A, Katul G, Law B, Lee XH, Malhi Y, Meyers T, Munger W, Oechel W, Paw KT, Pilegaard K, Schmid HP, Valentini R, Verma S, Vesala T, Wilson K, Wofsy S (2001) FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bull Am Meteorol Soc* 82(11):2415–2434
72. Friend AD, Arneth A, Kiang NY, Lomas M, Ogee J, Rodenbeck C, Running SW, Santaren JD, Sitch S, Viovy N, Woodward FI, Zaehle S (2007) FLUXNET and modelling the global carbon cycle. *Glob Change Biol* 13(3):610–633
73. Augustine JA, DeLuisi JJ, Long CN (2000) SURFRAD - a national surface radiation budget network for atmospheric research. *Bull Am Meteorol Soc* 81(10):2341–2357
74. Hari P, Andreae MO, Kabat P, Kulmala M (2009) A comprehensive network of measuring stations to monitor climate change. *Boreal Environ Res* 14(4):442–446
75. Jacquemoud S, Baret F (1990) Prospect - a model of leaf optical-properties spectra. *Remote Sens Environ* 34(2):75–91
76. Feret JB, Francois C, Asner GP, Gitelson AA, Martin RE, Bidet LPR, Ustin SL, le Maire G, Jacquemoud S (2008) PROSPECT-4 and 5: Advances in the leaf optical properties model separating photosynthetic pigments. *Remote Sens Environ* 112(6):3030–3043
77. Sims DA, Gamon JA (2002) Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. *Remote Sens Environ* 81(2–3):337–354
78. Martin G, Jossier SA, Bornman JF, Vogelmann TC (1989) Epidermal focusing and the light microenvironment within leaves of medicago-sativa. *Physiol Plant* 76(4):485–492
79. Farquhar GD, Caemmerer SV, Berry JA (1980) A biochemical-model of photosynthetic CO<sub>2</sub> assimilation in leaves of C-3 species. *Planta* 149(1):78–90
80. Tardieu F, Simonneau T (1998) Variability among species of stomatal control under fluctuating soil water status and evaporative demand: modelling isohydric and anisohydric behaviours. *J Exp Bot* 49:419–432
81. Emerson R (1929) The relation between maximum rate of photosynthesis and concentration of chlorophyll. *J Gen Physiol* 12(5):609–622
82. Peñuelas J, Baret F, Filella I (1995) Semiempirical indexes to assess carotenoids chlorophyll-a ratio from leaf spectral reflectance. *Photosynthetica* 31(2):221–230

83. Moya I, Camenen L, Evain S, Goulas Y, Cerovic ZG, Latouche G, Flexas J, Ounis A (2004) A new instrument for passive remote sensing 1, measurements of sunlight-induced chlorophyll fluorescence. *Remote Sens Environ* 91(2):186–197
84. Maxwell K, Johnson GN (2000) Chlorophyll fluorescence - a practical guide. *J Exp Bot* 51(345):659–668
85. Ross J (1981) The radiation regime and architecture of plant stands. Dr. W. Junk Publishers, The Hague
86. Myneni R, Ross JE (1991) Photon-vegetation interactions. Springer, Berlin, Heidelberg
87. Malenovsky Z, Mishra KB, Zemek F, Rascher U, Nedbal L (2009) Scientific and technical challenges in remote sensing of plant canopy reflectance and fluorescence. *J Exp Bot* 60(11):2987–3004
88. Monsi M, Saeki T (1953) Über den lichtfaktor in den pflanzen-gesellschaften und seine bedeutung für die stoffproduktion. *Jpn J Bot* 14:22–52
89. Monsi M, Saeki T (2005) On the factor light in plant communi-ties and its importance for matter production. *Ann Bot* 95(3):549–567
90. Lieffers V, Messier C, Stadt K, Gendron F, Comeau P (1999) Predicting and managing light in the understory of boreal forests. *Can J For Res* 29(6):796–811
91. Campbell GS, Norman JM (1998) Introduction to environmen-tal biophysics, 2nd edn. Springer, New York
92. Breda NJJ (Nov 2003) Ground-based measurements of leaf area index: a review of methods, instruments and current controversies. *J Exp Bot* 54:2403–2417
93. Jonckheere I, Fleck S, Nackaerts K, Muys B, Coppin P, Weiss M, Baret F (Jan 2004) Review of methods for in situ leaf area index determination - Part I theories, sensors and hemispherical photography. *Agric For Meteorol* 121:19–35
94. Ni WG, Li XW, Woodcock CE, Roujean JL, Davis RE (Dec 1997) Transmission of solar radiation in boreal conifer forests: Measurements and models. *J Geophys Res Atmos* 102:29555–29566
95. Möttus M, Sulev M (2006) Radiation fluxes and canopy trans-mittance: Models and measurements inside a willow canopy. *J Geophys Res Atmos* 111(D2):D02109
96. Chelle M (2006) Could plant leaves be treated as Lambertian surfaces in dense crop canopies to estimate light absorption? *Ecol Modell* 198(1–2):219–228
97. Asner GP, Wessman CA, Schimel DS, Archer S (Mar 1998) Variability in leaf and litter optical properties: Implications for BRDF model inversions using AVHRR, MODIS, and MISR. *Remote Sens Environ* 63:243–257
98. Baldocchi D, Collineau S (1994) The physical nature of solar radiation in heterogeneous canopies: spatial and temporal attributes. In: Exploitation of environmental heterogeneity by plants: ecophysiological processes above- and below ground. Academic, San Diego, pp 21–71
99. Reifsnyder W, Furnival G, Horowitz J (1971–1972) Spatial and temporal distribution of solar radiation beneath forest canopies. *Agric Meteorol* 9:21–37
100. DePury DGG, Farquhar GD (May 1997) Simple scaling of photosynthesis from leaves to canopies without the errors of big-leaf models. *Plant Cell Environ* 20:537–557
101. Oker-Blom P (1985) Photosynthesis of a Scots pine shoot – simulation of the irradiance distribution and photosynthesis of a shoot in different radiation-fields. *Agric For Meteorol* 34(1):31–40
102. Myneni RB, Asrar G, Wall GW, Kanemasu ET, Impens I (1986) Canopy architecture, irradiance distribution on leaf surfaces and consequent photosynthetic efficiencies in heteroge-neous plant canopies. 2 results and discussion. *Agric For Meteorol* 37(3):205–218
103. Stenberg P (1998) Implications of shoot structure on the rate of photosynthesis at different levels in a coniferous canopy using a model incorporating grouping and penumbra. *Funct Ecol* 12(1):82–91
104. Gu LH, Baldocchi D, Verma SB, Black TA, Vesala T, Falge EM, Dowty PR (2002) Advantages of diffuse radiation for terrestrial ecosystem productivity. *J Geophys Res Atmos* 107(D5–6):4050
105. Roderick ML, Farquhar GD, Berry SL, Noble IR (Sept 2001) On the direct effect of clouds and atmospheric particles on the pro-ductivity and structure of vegetation. *Oecologia* 129:21–30
106. Gu LH, Baldocchi DD, Wofsy SC, Munger JW, Michalsky JJ, Urbanski SP, Boden TA (2003) Response of a deciduous forest to the Mount Pinatubo eruption: Enhanced photosynthesis. *Science* 299(5615):2035–2038
107. Mercado LM, Bellouin N, Sitch S, Boucher O, Huntingford C, Wild M, Cox PM (Apr 2009) Impact of changes in diffuse radia-tion on the global land carbon sink. *Nature* 458:1014–1017
108. Vesala T, Markkanen T, Palva L, Siivola E, Palmroth S, Hari P (Feb. 2000) Effect of variations of PAR on CO<sub>2</sub> exchange estimation for Scots pine. *Agric For Meteorol* 100:337–347
109. Schmid HP (1997) Experimental design for flux measure-ments: matching scales of observations and fluxes. *Agric For Meteorol* 87(2–3):179–200
110. Schaepman-Strub G, Schaepman ME, Painter TH, Dangel S, Martonchik JV (2006) Reflectance quantities in optical remote sensing-definitions and case studies. *Remote Sens Environ* 103(1):27–42
111. Nicodemus FE (1970) Reflectance nomenclature and direc-tional reflectance and emissivity. *Appl Opt* 9(6):1474–1475
112. Baret F, Guyot G (1991) Potentials and limits of vegetation indexes for LAI and APAR assessment. *Remote Sens Environ* 35(2–3):161–173
113. McCallum I, Wagner W, Schmulilius C, Shvidenko A, Obersteiner M, Fritz S, Nilsson S (2010) Comparison of four global FAPAR datasets over Northern Eurasia for the year 2000. *Remote Sens Environ* 114(5):941–949
114. Asrar G (1989) Theory and applications of optical remote sensing. Wiley, New York
115. Begue A, Desprat JF, Imbernon J, Baret F (1991) Radiation use efficiency of pearl-millet in the Sahelian zone. *Agric For Meteorol* 56(1–2):93–110
116. Gobron N, Pinty B, Aussenat O, Chen JM, Cohen WB, Fensholt R, Gond V, Huemmrich KF, Lavergne T, Melin F,



- Privette JL, Sandholt I, Taberner M, Turner DP, Verstraete MM, Widlowski JL (2006) Evaluation of fraction of absorbed photosynthetically active radiation products for different canopy radiation transfer regimes: methodology and results using joint research center products derived from SeaWiFS against ground-based estimations. *J Geophys Res Atmos* 111(D13110):1–15
117. Russel G, Jarvis P, Monteith J (1989) Absorption of radiation by canopies and stand growth. In: Russel G, Marshall B, Jarvis PG (eds) *Plant canopies: their growth, form and function*. Cambridge University Press, New York, pp 21–39
118. Palva L, Garam E, Manoochehri F, Sepponen R, Hari P, Rajala K, Ruotoistenmaki H, Seppala I (1998) A novel multipoint measuring system of photosynthetically active radiation. *Agric For Meteorol* 89(2):141–147
119. Palva L, Markkanen T, Siivola E, Garam E, Linnavuo M, Nevas S, Manoochehri F, Palmroth S, Rajala K, Ruotoistenmaki H, Vuorivirta T, Seppala I, Vesala T, Hari P, Sepponen R (2001) Tree scale distributed multipoint measuring system of photosynthetically active radiation. *Agric For Meteorol* 106(1):71–80
120. Baret F, Hagolle O, Geiger B, Bicheron P, Miras B, Huc M, Berthelot B, Nino F, Weiss M, Samain O, Roujean JL, Leroy M (2007) LAI, fAPAR and fCover CYCLOPES global products derived from VEGETATION - part 1: Principles of the algorithm. *Remote Sens Environ* 110(3):275–286
121. Chen JM (1996) Optically-based methods for measuring seasonal variation of leaf area index in boreal conifer stands. *Agric For Meteorol* 80(2–4):135–163
122. Nilson T (1971) Theoretical analysis of frequency of gaps in plant stands. *Agric Meteorol* 8(1):25–38
123. Lang ARG (1986) Leaf-area and average leaf angle from transmission of direct sunlight. *Aust J Bot* 34(3):349–355
124. Welles JM, Norman JM (1991) Instrument for indirect measurement of canopy architecture. *Agron J* 83:818–825
125. Baret F, Andrieu B, Folmer J, Hanocq J, Sarrouy C (1993) Gap fraction measurement from hemispherical infrared photography and its use to evaluate PAR interception efficiency. In: *Crop structure and microclimate: characterization and applications*, INRA edn. Paris, pp 359–372
126. Leblanc SG, Chen JM, Fernandes R, Deering DW, Conley A (2005) Methodology comparison for canopy structure parameters extraction from digital hemispherical photography in boreal forests. *Agric For Meteorol* 129:187–207
127. Cescatti A (2007) Indirect estimates of canopy gap fraction based on the linear conversion of hemispherical photographs: Methodology and comparison with standard thresholding techniques. *Agric For Meteorol* 143:1–12
128. Lang M, Kuusk A, Möttus M, Rautiainen M, Nilson T (2010) Canopy gap fraction estimation from digital hemispherical images using sky radiance models and a linear conversion method. *Agric For Meteorol* 150(1):20–29
129. Campbell GS (1986) Extinction coefficients for radiation in plant canopies calculated using an ellipsoidal inclination angle distribution. *Agric For Meteorol* 36(4):317–321
130. Asrar G, Fuchs M, Kanemasu ET, Hatfield JL (1984) Estimating absorbed photosynthetic radiation and leaf-area index from spectral reflectance in wheat. *Agron J* 76(2):300–306
131. Myneni RB, Williams DL (1994) On the relationship between fAPAR and NDVI. *Remote Sens Environ* 49(3):200–211
132. Weiss M, Baret F (1999) Evaluation of canopy biophysical variable retrieval performances from the accumulation of large swath satellite data. *Remote Sens Environ* 70(3):293–306
133. Weiss M, Baret F, Myneni RB, Pragnere A, Knyazikhin Y (2000) Investigation of a model inversion technique to estimate canopy biophysical variables from spectral and directional reflectance data. *Agronomie* 20(1):3–22
134. Roujean JL, Breon FM (1995) Estimating PAR absorbed by vegetation from bidirectional reflectance measurements. *Remote Sens Environ* 51(3):375–384
135. Myneni RB, Hoffman S, Knyazikhin Y, Privette JL, Glassy J, Tian Y, Wang Y, Song X, Zhang Y, Smith GR, Lotsch A, Friedl M, Morisette JT, Votava P, Nemani RR, Running SW (2002) Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data. *Remote Sens Environ* 83(1–2):214–231
136. Bacour C, Baret F, Beal D, Weiss M, Pavageau K (2006) Neural network estimation of LAI, fAPAR, fCover and LAIxC(ab), from top of canopy MERIS reflectance data: Principles and validation. *Remote Sens Environ* 105(4):313–325
137. Fensholt R, Sandholt I, Rasmussen MS (2004) Evaluation of MODIS LAI, fAPAR and the relation between fAPAR and NDVI in a semi-arid environment using in situ measurements. *Remote Sens Environ* 91(3–4):490–507
138. Morisette JT, Baret F, Privette JL, Myneni RB, Nickeson JE, Garrigues S, Shabanov NV, Weiss M, Fernandes RA, Leblanc SG, Kalacska M, Sanchez-Azofeifa GA, Chubey M, Rivard B, Stenberg P, Rautiainen M, Voipio P, Manninen T, Piliant AN, Lewis TE, Iames JS, Colombo R, Meroni M, Busetto L, Cohen WB, Turner DP, Warner ED, Petersen GW, Seufert G, Cook R (2006) Validation of global moderate-resolution LAI products: A framework proposed within the CEOS Land Product Validation subgroup. *IEEE Trans Geosci Remote Sens* 44(7):1804–1817
139. Weiss M, Baret F, Garrigues S, Lacaze R (2007) LAI and fAPAR CYCLOPES global products derived from VEGETATION, part 2: validation and comparison with MODIS collection 4 products. *Remote Sens Environ* 110(3):317–331
140. Wiegand CL, Maas SJ, Aase JK, Hatfield JL, Pinter PJ, Jackson RD, Kanemasu ET, Lapitan RL (1992) Multisite analyses of spectral-biophysical data for wheat. *Remote Sens Environ* 42(1):1–21
141. Sellers PJ (1985) Canopy reflectance, photosynthesis and transpiration. *Int J Remote Sens* 6(8):1335–1372
142. Baret F, Pavageau K, Béal D, Weiss M, Berthelot B, Regner P (2006) Algorithm theoretical basis document for MERIS top of atmosphere land products (TOA\_VEG). INRA-CSE, 2006
143. Deng F, Chen JM, Plummer S, Chen MZ, Pisek J (2006) Algorithm for global leaf area index retrieval using satellite imagery. *IEEE Trans Geosci Remote Sens* 44(8):2219–2229

144. Baret F, Weiss M, Lacaze R, Camacho F, Pacholczyk P, Makhmara H, Smets B (2010) Consistent and accurate LAI, FAPAR and FCOVER global products: principles and evaluation of GEOV1 products. In Sobrino (ed) Proceedings of the third international symposium on recent advances in quantitative remote sensing, Valencia (Spain), pp 208–213

## Photovoltaic Energy, Introduction

DANIEL LINCOT

Institute of Research and Development of Photovoltaic Energy, Chatou, Cedex, France

### Article Outline

Glossary

### Glossary

**III-V Solar cells** Solar cells based on compound combining elements from the Ga and As columns (III and V).

**Cadmium telluride (CdTe) solar cells** Solar cells based on this compound, used in the form of thin films.

**Chalcopyrite solar cells** Solar cells based on the compound  $\text{Cu}(\text{In,Ga})\text{Se}_2$ , also noted CIGS, in the form of thin films.

**Dye sensitized solar cells (DSSC)** Solar cells based on mesoscopic titanium oxide thin film sensitized with dye photoactive molecules and impregnated by an electrolyte.

**Hot carrier solar cell** New high efficiency concept allowing to convert high energy photons in electrical charges in the external circuit without thermal losses.

**Life cycle analysis (LCA)** To quantify all the steps from mining, utilization to recycling in terms of energy consumption, material utilization, environmental and health impacts.

**Multijunction solar cell** High efficiency solar cell based on the association of several elementary solar cells made of sing junctions.

**Organic solar cells** Solar cells based on organic components like carbon fullerenes and polymers, blended in the form of thin films.

**Pay back time** Time needed by a solar cell under operation to reimburse the total energy used for its fabrication.

**Photovoltaics** Conversion of photon energy to electricity.

**Silicon solar cells** Solar cells based on silicon element, either in crystalline or amorphous forms.

**Solar cells** Device allowing to absorb photon energy and convert it to electricity in an external circuit.

**Up, down conversion** New high efficiency concept using optical processes allowing to convert low (resp. high) energy photons to medium visible energy photons for maximum conversion efficiency.

Photovoltaics is the direct conversion of solar energy into electricity. It results from the fundamental mechanism of absorption of photons in matter, with the excitation of electrons from their equilibrium lower energy state to a nonequilibrium excited state of higher energy. That means that electrons are being transferred to more negative electrical potential. Then, they usually return to equilibrium by giving back the initial photon energy in form of thermal energy (with the interactions with phonons), light with the emission of new photons via luminescence processes or chemical species via electrochemical oxydo reduction processes in the case of photosynthesis. The uniqueness and beauty of photovoltaics is to “plug” on the initial step when electrons are just excited to a lower potential, and to have them directly transferred in an external circuit where the energy can be used directly in the electrical form. The device to do it is just a solar cell. However, to have it efficient imposes to be able to compete with the naturally occurring spontaneous processes! This was not easy and from the discovery of the photovoltaic effect in 1839 by Edmond Becquerel to the first efficient silicon solar cell in 1954 it took more than one century and then 50 years more to reach the years 2000s to assist to the large scale industrial endeavor of photovoltaic conversion of solar energy, bringing for the first time in the human history this new renewable energy technology as an alternative to fossil fuels and nuclear utilizations. While laboratory record efficiency for any photovoltaic cells is reaching the incredible value of 43%, approaching the 50% level, more than 20 GW of photovoltaic peak power sources have been produced by the industry in 2010. This is the

reason why this section of the encyclopedia on photovoltaics is timely.

It aims to give the most advanced analyses of the context of photovoltaic deployment in the present days and the frontiers of key solar cell technologies and concepts in their booming research environment. These entries will be authored by upmost specialists in each field, who have been or are still directly active in key advances.

► [PV Policies and Markets](#), authored by Dr. Wolfgang Palz, former European Union Official in charge of photovoltaic R&D programs for long time, and presently President of the World Council on Renewable Energies, will present an insight analysis of photovoltaic policy and markets, discuss specific questions associated with fast changes of players and fast decrease of prices and presents his views on challenges for the future.

► [Photovoltaics, status of](#) presents an extensive review of the status of photovoltaics, starting with an overview of the existing and emerging technologies, followed by the analysis of research and developments issues and presenting then the state of the art of the photovoltaic industry. In the second part of the entry, it is presented the photovoltaic market with specific focuses on the European Union, the Asia and Pacific region, and then North America. This broad view entry is authored by Dr Arnulf Jaeger Waldau, from the Joint Research Center of the European Commission. Dr Arnulf Jaeger Waldau is the author of the famous series of yearly edited PV Status reports.

Life cycle analysis (LCA) is a key concern in the photovoltaic domain in order to assess and quantify their benefit for contributing to creating a fully sustainable society. Life cycle analysis is thus a very important emerging domain also in all other area of the human activity, considering for a given product all its aspects from mining, processing, use, and recycling steps. Prof. Vasilis Fthenakis, from the Brookhaven National Laboratory, is a pioneer and one of the most eminent researcher on LCA in the field of photovoltaics. He authored this specific area in entry ► [Solar Cells: Energy Payback Times and Environmental Issues](#) by considering two aspects, one is the energy payback time of solar cells, that means the time needed for the device to reimburse by its own production the energy which has been used for its production, the

other one concerns environmental issues. The analysis and discussion will cover the different photovoltaic technologies at their present state of the art and give detailed information for the reader.

After these three supplementary entries, devoted to economical and environmental aspects of photovoltaics, the next entries will focus on each photovoltaic technology, ranging of established crystalline silicon technologies and emerging thin film technologies to very high efficiency devices and new concepts. Organic and hybrid photovoltaic devices will complete the panorama of actual and future photovoltaic technologies.

The entry on crystalline wafer-based silicon technologies, ► [Silicon Solar Cells, Crystalline](#), which is by far, the majoritary technology (87% of the market in 2010) is authored by a group of researchers specialists of these technologies under the lead of Prof. Santo Martinuzzi from the University of Marseille, a pioneer of silicon photovoltaics, well-known for his contributions to the characterization and understanding of electrical defects and passivation effects in silicon material. The entry will recall some historical aspects of the crystalline silicon technology and fundamental processes used for silicon production, purification, and device elaboration. It will present very important phenomena associated to defect characterization and passivation which control the cell efficiencies. It will be concluded by presenting the remarkable developments in new solar cell architectures which pave the way of future generations of crystalline silicon solar cells.

► [Silicon Solar Cells, Thin-film](#) is authored by Prof. Christopher Wronsky, from Pennsylvania State University and Prof. Nicolas Wyrsh from the Ecole Polytechnique Fédérale de Lausanne (EPFL) and devoted to thin film silicon solar cells, based on long standing amorphous to newly introduced microcrystalline materials, and representing 5% of the market in 2010. C.W. is a pioneer of amorphous silicon solar cells, which is the historical thin film solar cell technology, and specialist of their fundamental properties. He discovered key physical phenomenon known as the Staebler-Wronski effect. N.W. is highly involved in the unique R&D experience in microcrystalline solar cells alone or combined with amorphous solar cells known as the micromorph concept. The entry will go in depth along the whole sequences of thin film Si solar

cells, from the growth to devices physics and properties. It contains unique descriptions of physicochemical properties of layers controlling the device efficiencies and properties. Future directions in this promising technology are addressed at the last section of the entry.

► [CdTe Solar Cells](#) deals with another thin film technology based on cadmium telluride devices which has experienced an exceptional development since 2005, moving from a secondary industrial position to the first one in 2009, only in 5 years. It represents 5% of the market in 2010. CdTe technology is now the leading thin film technology and was capable to demonstrate in practice for the first time the cost reduction breakthrough expected for thin film technologies. The entry is written by Prof. Ayodya Tiwari and S. Buecheler, L. Kranz, J. Perrenoud from his group. Prof. Tiwari is a leading researcher in the field of CdTe solar cells. The entry will bring the reader from the description of state of the art of CdTe solar cells to in depth presentation and analysis of key processing steps and properties of layers and devices. Environmental and material resource aspects which are real concerns for this technology will be also considered, together with future directions.

► [Solar Cells, Chalcopyrite-based Thin Film](#) will continue this travel within thin film solar cell technologies with an another exceptional nonsilicon material based on the ternary compound  $\text{CuInSe}_2$  noted CIS and its alloys with Ga, Al, S (mainly with Ga noted CIGS). CIS technologies are just starting their industrial endeavor (1.6%) while cell record efficiencies at the laboratory level now overpasses 20%, similar to that obtained with wafer-based polycrystalline silicon solar cells. This opens a new direction for thin film solar cells in competing also in the field of high efficiency devices. Prof. Hans Schock for Helmholtz Zentrum Berlin (HZB), a pioneer and main actor of the CIGS adventure, recipient of the famous Becquerel price, has authored this entry with insights in key aspects of growth and properties of CIGS layers, junction formation, and device properties, from fundamental to process oriented focus. He will discuss present bottlenecks and future challenges of this technology.

The two next entries are devoted to the field of emerging technologies based on organic solar cells. This represents a new avenue for photovoltaics which was historically based on inorganic materials like

silicon and compound semiconductors. The idea is that using organic materials may allow new cost reduction breakdowns and the use of low temperature processing compatible with plastic substrates. Up to the 90s, the use of organic photoactive layers was hampered by limited performances in charge separation and electronic transport required for efficient devices. A revolutionary concept was demonstrated in 91 in the case of hybrid solar cells is that of interpenetrated networks junctions, and then also introduced in the field of full organic junctions, and named bulk heterojunction concept. Since this time, organic-based solar cells are associated to a booming R&D activity.

► [Mesoscopic Solar Cells](#) presents the advent of hybrid nanostructured solar cells, based on the sensitization of nanoporous oxide layers, mostly titanium oxide, with dye molecules, followed by the impregnation by a liquid electrolyte containing an energetically suited redox couple (mostly iodide-iodide). These cells, also named Dye Sensitized Solar Cells (DSSC) which are photoelectrochemical cells, reach impressive performances, up to 12% record cells and start to be industrially produced for niche markets. The author of the entry “► [Mesoscopic Solar Cells](#)” is Prof. Michael Graetzel, from the Ecole Polytechnique Fédérale de Lausanne. He is the inventor of DSSCs in 91 and since this time leads with his group key advances in the field related to the optimization of the electrode, the dye molecules, and also the replacement of the liquid electrolyte by a solid state one. His entry will present key aspects of this technology and future challenges.

► [Organic Solar Cells](#) is devoted to the field of full organic solar cells, based on carbon fullerenes, polymers, and small organic molecules. Will these “plastic solar cells,” as their sister devices of organic electroluminescence devices (OLED), bring a new revolution in photovoltaics, together with Dye cells? This is clearly an open avenue which can be anticipated from the constant progress in performances and stability during the last period, with record cell efficiencies increasing from a few per cent in 2000 to about 10% now. Even if the module efficiencies are still about a few per cent and time stability is a concern, niche markets are already open. This status is the result of key researchers in the field as Prof. Saricifti of the Linz

Institute for Organic Solar Cells and Prof. Ching Tang from the University of Rochester who are authoring with Dr Daniel Glowacki this entry on organic solar cells. From operating principles to actual state and future directions, key aspects of this booming field will be presented and analyzed in great detail.

► **Solar Cells, High-Efficiency** opens the magic box of photovoltaics with the graal of obtaining ultra high Efficiency devices. While the efficiency of standard single junctions is theoretically limited to around 32%, according to the Shockley Queisser (S-Q) theory, and being approached by record silicon devices (25%) and even closer recently with GaAs devices (28%), theoretical studies predict that the efficiency could be as high as around 85%. Increasing the efficiency of solar cells beyond the S-Q limit is thus a key challenge for photovoltaic research, with achieving for instance 50% efficiencies as a practical achievable goal. In fact devices based on multijunctions, which exist for long time, are already able to escape the S-Q limit. A fascinating race is engaged between different laboratories using triple junctions based on gallium arsenide (GaAs) and it allows with P and Al operating under concentration, and recent record is now 43.5%. With about 30 layers deposited epitaxially, these devices are the cathedrals of photovoltaic technology which may gain 1 or to levels to reach 50%. Beyond the multijunction concept, new revolutionary concepts base on simpler device architectures have been proposed in the last decade, based on changing the wavelengths of the photons by up and down conversion optical processes (photon conversion concept), or by introducing additional levels in the band gap of single junctions (intermediate gap), or by collecting the generated charges just after their formation before partial relaxation (hot carrier concept). This entry will be authored by Dr Jean François Guillemoles from IRDEP and Dr Alex Freundlich from the University of Houston who are eminent specialists in these fields.

The entry will thus give a detailed view of the state of the art of photovoltaics with a specific focus on solar cell technologies which are the active component, the engine, in energy transformation, and a booming domain. Other aspects, which have not been treated yet in the section, but which are also very important are photovoltaic systems and PV integration in architecture or in solar farms.

## Photovoltaics, Status of

ARNULF JÄGER-WALDAU

European Commission, DG Joint Research Centre,  
Institute for Energy and Transport;  
Renewable Energy Unit, Ispra (VA), Italy

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Technology Overview  
Research and Development in Photovoltaics  
The Photovoltaic Industry  
The Photovoltaic Market  
Conclusions  
Future Directions  
Exchange Rates  
Bibliography

### Glossary

**Photovoltaics (PV)** PV is a method of generating electrical power by converting solar radiation into direct current electricity using semiconductors that exhibit the photovoltaic effect and are called solar cells.

**Photovoltaic capacity** The capacity of PV systems is given in  $W_p$  (watt peak). This characterizes the maximum DC (direct current) output of a solar module under standard test conditions, that is, at a solar radiation of  $1,000 \text{ W/m}^2$  and at a temperature of  $25^\circ\text{C}$ .

**Photovoltaic electricity generation** The actual electricity generation potential of a photovoltaic electricity system depends on the solar radiation and the system performance, which depends on the BOS component losses. For a solar radiation between 600 and  $2,200 \text{ kWh/m}^2/\text{year}$  an average PV system can produce between 450 and  $1,650 \text{ kWh}$  of AC electricity.

**Photovoltaic module and photovoltaic system** A number of solar cells form a solar “Module” or “Panel,” which can then be combined to solar systems, ranging from a few Watts of electricity output to multimegawatt power stations.

**Photovoltaic (PV) energy system** A PV system is composed of three subsystems:

- On the power generation side, a subsystem of PV devices (cells, modules, arrays) converts sunlight to direct current (DC) electricity.
- On the power use side, the subsystem consists mainly of the load, which is the application of the PV electricity.
- Between these two, we need a third subsystem that enables the PV-generated electricity to be properly applied to the load. This third subsystem is often called the “balance of system” or BOS.

**Polisilicon or polycrystalline silicon** Polisilicon or Polycrystalline silicon is a material consisting of small silicon crystals.

**EC framework program** This is the main instrument of the European Union for funding research.

**Feed-in tariff** A feed-in tariff is a policy mechanism which obliges regional or national electric grid utilities to buy renewable electricity (electricity generated from renewable sources, such as solar power, wind power, wave and tidal power, biomass, hydro-power, and geothermal power) from all eligible participants at a fixed price over a fixed period of time.

**Power purchase agreement (PPA)** A PPA is a legal contract between an electricity generator (provider) and a power purchaser (host).

**Solar cell production capacities**

- In the case of wafer silicon-based solar cells, only the cells
- In the case of thin films, the complete integrated module
- Only those companies which actually produce the active circuit (solar cell) are counted
- Companies which purchase these circuits and make cells are not counted.

### Definition of the Subject

Solar energy is the most abundant of all energy resources, and the rate at which solar energy is intercepted by the Earth is about 10,000 times greater than the rate at which all energy is used on this planet.

Solar energy can be used by a family of technologies capable of being integrated amongst themselves, as well as with other renewable energy technologies. The solar technologies can deliver heat, cooling, electricity, lighting, and fuels for a host of applications.

The conversion of solar energy into electricity, the photovoltaic effect, was discovered by Alexandre-Edmond Becquerel in 1839. However, it took more than a 100 years, until in 1954 scientists at the Bell Laboratories unveiled the first modern solar cell, using a silicon semiconductor to convert light into electricity.

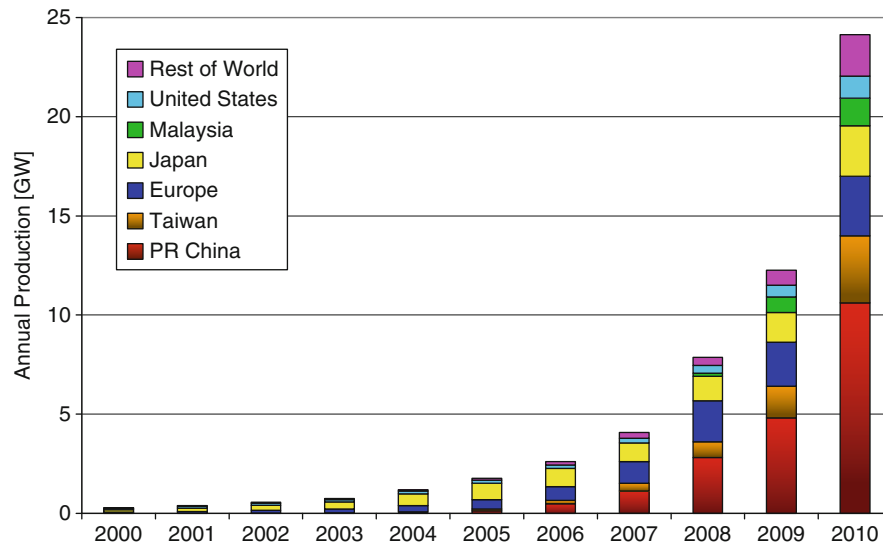
With the oil crisis of the 1970s, many countries in the world started solar energy research and development (R&D) programs, but it took another 20 years until the first market implementation programs for grid-connected solar photovoltaic electricity generation systems started in the 1990s and began to prepare the basis for the development of a photovoltaics industry.

### Introduction

For more than 10 years, photovoltaics has been one of the fastest growing industries with growth rates well beyond 40% per annum. This growth is driven not only by the progress in materials and processing technology, but by market introduction programs in many countries around the world and the increased volatility and mounting fossil energy prices. Despite the negative impacts of the economic crisis which started in 2008, photovoltaics is still growing at an extraordinary pace.

Reported production data for the global cell production in 2010 vary between 18 and 27 GW. The significant uncertainty in the data for 2010 is due to the highly competitive market environment, as well as the fact that some companies report shipment figures, others report sales and again others report production figures. In addition, the difficult economic conditions and increased competition led to a decreased willingness to report confidential company data. The data presented, collected from various companies and colleagues were compared to various data sources and thus led to an estimate of 24 GW (Fig. 1), representing a doubling of production compared to 2009.

Since 2000, total PV production increased almost by two orders of magnitude, with annual growth rates



**Photovoltaics, Status of. Figure 1**

Worldwide PV Production from 2000 to 2010 (Data Source: Navigant Consulting [1, 2], PV News [3], Photon International [4], and own analysis)

between 40% and 90%. The most rapid growth in annual production over the last 5 years could be observed in Asia, where China and Taiwan together now account for almost 60% of worldwide production.

Public-traded companies manufacturing solar products, or offering related services, have attracted a growing number of private and institutional investors. In 2010, worldwide new investments into the renewable energy and energy efficiency sectors increased to a new record of \$243 billion (€ 187 billion), up 30% from 2009 and for the third year in a row solar power attracted, behind wind, the second largest amount of new investments into renewable energies [5]. Europe was still the leading region in terms of renewable energy investments, totaling \$94.4 billion (€ 72.6 billion), followed by Asia/Oceania with \$82.8 billion (€ 63.7 billion) and the Americas with \$65.8 billion (€ 50.6 billion) [6].

The number of consulting companies and financial institutions offering market studies and investment opportunities has considerably increased in the last years, and business analysts are very confident that despite the current economic turmoil, the Photovoltaics sector is in a healthy long-term condition. Following the stock market decline, as a result of the financial turmoil, the PPVX (Photon Photovoltaic stock index)

declined to 2,095 points at the end of 2008. At the beginning of July 2011, the index stood at 2,107 points and the market capitalization of the 30-PPVX companies was € 36.4 billion.

The PPVX is a noncommercial financial index published by the solar magazine “Photon” and “Öko-Invest.” The index started on August 1, 2001 with 1,000 points and 11 companies and is calculated weekly using the Euro as reference currency. Only companies which made more than 50% of their sales in the previous year with PV products or services are included [7]. Please note that the composition of the index changes as new companies are added and others have to leave the index.

At the end of 2010, about 48% or \$94.8 billion (€ 72.9 billion) of the \$194.3 billion (€149.5 billion) global “green stimulus” money from governments aimed to help relieve the effect of the recession, had reached the markets [5]. For 2011 another \$68 billion (€52.3 billion) are expected.

Market predictions for the 2011 PV market vary between 17.3 GW by the Navigant Consulting conservative scenario [2], 22 GW by IMS Research [8], and 24.9 GW by iSuppli [9], with a consensus value in the 20 GW range. Massive capacity increases are underway or announced and if all of them are realized, the

worldwide production capacity for solar cells would exceed 80 GW at the end of 2012. This indicates that even with the optimistic market growth expectations, the planned capacity increases are way above the market growth. The consequence would be the continuation of the low utilization rates and therefore a continued price pressure in an over-supplied market. Such a development will accelerate the consolidation of the photovoltaics industry and spur more mergers and acquisitions.

The current solar cell technologies are well established and provide a reliable product, with sufficient efficiency and energy output for at least 25 years of lifetime. This reliability, the increasing potential of electricity interruption from grid overloads, as well as the rise of electricity prices from conventional energy sources, add to the attractiveness of Photovoltaic systems.

About 80% of the current production uses wafer-based crystalline silicon technology. A major advantage of this technology is that complete production lines can be bought, installed, and be up and producing within a relatively short time frame. This predictable production start-up scenario constitutes a low-risk placement with calculable return on investments. However, the temporary shortage in silicon feedstock and the market entry of companies offering turn-key production lines for thin-film solar cells led to a massive expansion of investments into thin-film capacities between 2005 and 2009. More than 200 companies are involved in the thin-film solar cell production process ranging from R&D activities to major manufacturing plants.

Projected silicon production capacities available for solar in 2012 vary between 140,000 t from established polysilicon producers, up to 185,000 t including the new producers [10] and 250,000 t [11]. The possible solar cell production will in addition depend on the material use per Wp. Material consumption could decrease from the current 8 g/Wp to 7 g/Wp or even 6 g/Wp, but this might not be achieved by all manufacturers.

Similar to other technology areas, new products will enter the market, enabling further cost reduction. Concentrating Photovoltaics (CPV) is an emerging market with approximately 30–40 MW cumulative installed capacity at the end of 2010. In addition, Dye-cells are getting ready to enter the market as well.

The growth of these technologies is accelerated by the positive development of the PV market as a whole.

Further photovoltaic system cost reductions will depend not only on the technology improvements and scale-up benefits in solar cell and module production, but also on the ability to decrease the system component costs as well as the whole installation, projecting, operation, and financing costs.

## Technology Overview

In order to give an overview of the technologies discussed, this entry will list the different technologies under investigation and in production with general descriptions rather than techno-scientific details, which are better championed by respective specialists. For categorization reasons, the solar cell technologies are divided into two categories of “existing PV technologies” and “emerging technologies.”

### Existing PV Technologies

These technologies include wafer-based crystalline silicon PV as well as the thin-film technologies *copper-indium/gallium-disulfide/diselenide* (CIGS), *cadmium telluride* (CdTe), and thin-film silicon PV (amorphous and microcrystalline). Mono- and poly(multi)crystalline silicon PV (including ribbon technologies) are the dominant technologies on the PV market, with a 2010 market share of about 80%.

*Crystalline silicon* modules are typically produced in a processing sequence along a value chain that starts with purified silicon, which is melted and solidified using different techniques to produce ingots or ribbons with variable degrees of crystal perfection. The ingots are then shaped into bricks and sliced into thin wafers by wire-sawing. In the case of ribbons, wafers are cut from the sheet typically using a laser. Cut wafers and ribbons are processed into solar cells and interconnected in weatherproof packages designed to last for at least 25 years.

In the laboratory, the externally verified record cell conversion efficiency is 25.0% for *monocrystalline silicon* and 20.4% for *multicrystalline cells* [12–14] under standard reporting conditions (i.e., 1,000 W/m<sup>2</sup>, AM1.5, 25°C). The theoretical Shockley-Queisser limit of a single-junction Si solar cell is 31% conversion efficiency [15] while the specific maximum



efficiency for crystalline silicon has been calculated to be 29% [16].

Several variations for higher efficiency have been developed, for example, *heterojunction solar cells* and *interdigitated back contact solar cells*. These solar cells consist of a crystalline silicon wafer base with a (deposited) amorphous silicon emitter. The highest efficiency of heterojunction solar cells is 23% for a 100-cm<sup>2</sup> cell [17]. In an interdigitated back contact solar cell both the base and the emitter are contacted at the back of the cell, with as one of the advantages the avoidance of shading of the front of the cell by a top electrode. The highest efficiency of such a monocrystalline silicon back contact cell reported is 23.4% [18].

Wafers have decreased in thickness from 400  $\mu\text{m}$  in 1990 to less than 200  $\mu\text{m}$  in 2009 and have increased in area from 100 cm<sup>2</sup> to over 200 cm<sup>2</sup>. Modules have increased in efficiency from about 10% in 1990 to typically 13–15% today, with the best performers above 17%. In addition, manufacturing facilities have increased from the typical 1–5 MWp annual outputs in 1990 to hundreds of MWp for today's largest factories. The processes in the value chain have progressed significantly during recent years, but they still have potential for further large improvements. Commercial module efficiencies for wafer-based silicon PV range from 12% to 20%.

*Commercial thin-film technologies* include a range of material systems, from silicon-based cells to the chalcopyrite material family like cadmium telluride (CdTe) or copper indium gallium diselenide (CIGSS). These solar cells have a base material of a few micrometer thickness deposited on glass, metal, or plastic substrates with sizes up to 5.7 m<sup>2</sup>.

The *amorphous silicon* (a-Si) solar cell, introduced in 1976 [19] with initial efficiencies of 1–2%, has been the first commercially successful thin-film solar cell technology. Amorphous Si is a quasi-direct-bandgap material and hence has a high light absorption coefficient; therefore, the thickness of an a-Si cell can be 1,000 times thinner than that of a crystalline Si (c-Si) cell. Developing better efficiencies for a-Si has been limited by inherent material quality and by light-induced degradation – the Staebler–Wronski effect [20]. However, research efforts have successfully lowered the impact of the Staebler–Wronski effect to

around 10% or less by controlling the microstructure of the film. The highest stabilized efficiency reported is 10.1% [21].

Higher efficiency has been achieved by using multijunction technologies with alloy materials, for example, germanium and carbon to form semiconductors with lower or higher bandgaps, respectively, to cover a wider range of the solar spectrum [22]. Another approach to increase the efficiency of thin-film silicon devices is through a tandem consisting of a microcrystalline silicon bottom cell with an amorphous silicon top cell [23, 24]. Stabilized efficiencies of 12–13% have been measured for various laboratory devices [12].

*CdTe solar cells* using a heterojunction with CdS have shown significant promise, because CdTe has a suitable energy bandgap of 1.45 electron-volts (eV) with a high coefficient of light absorption. The best efficiency of this cell is 16.5% [25], and commercially available modules have an efficiency of around 10%. Goncalves et al. predicted that the maximum efficiency will be 17.6%, and future improvements will focus on how to further reduce manufacturing costs, which are already the lowest in the industry [26]. The toxicity of metallic cadmium and the relative scarcity of tellurium are issues commonly associated with this technology. CdTe itself is a semiconductor and only limited toxicological data are available. Therefore, the evaluation of potential health risks is so far based on other forms of cadmium [27]. The currently known toxic health effects described on a typical material safety data sheet are limited to dust inhalation and ingestion. Recent investigations by Zayed et al. on the acute oral and inhalation toxicity in rats show that the toxicity potential is much lower than that of cadmium [28]. But this potential hazard is mitigated by using a glass-sandwiched module design and by recycling the entire module, as well as industrial waste [27]. Contrary to the commonly associated scarcity of tellurium, Wadia et al. found that the currently known economic tellurium reserves would allow the installation of approximately 10 TW of CdTe solar cells [29].

The *copper indium gallium sulfur selenide* material family is the base for the highest efficiency thin-film solar cells so far. The CuInSe<sub>2</sub>/CdS solar cell was invented in the early 1970s at Bell Labs [30]. Incorporating Ga and/or S to produce CuInGa(Se,S)<sub>2</sub> (CIGSS)

results in the benefit of a widened bandgap depending on the composition [31]. CIGS-based solar cells yield a maximum efficiency of 20.3% [32], using a double graded layer of Ga in the absorption layer to realize both high current density and high open-circuit voltage. Due to higher efficiencies and lower manufacturing energy consumptions, CIGSS cells are currently in the industrialization phase with best commercial module efficiencies of up to 13.1% [33] for  $\text{CuInGaSe}_2$  and 8.6% for  $\text{CuInS}_2$  [34]. Contrary to the commonly associated scarcity of indium, Wadia et al. found that the currently known economic indium reserves would allow the installation of more than 10 TW of CIGSS based photovoltaic systems [29].

*High-efficiency solar cells* based on GaAs and InGaP (i.e., III-V semiconductors) have superior efficiencies but are also expensive, devices. Double- and triple-junction devices are currently being commercialized. An economically feasible application is the use of these cells in concentrator PV systems [35]. The most commonly used cell is a three-junction device based on GaInP/GaAs/Ge, with a record efficiency of 41.6% [36] under  $364\times$  concentrated light. Concentrator application requires a high fraction of direct (vs diffuse) irradiation, and is thus only suited for the sunbelt regions, that is, using the optical systems available so far.

### Emerging Technologies

These are technologies that are still under development and in laboratory or (pre-) pilot stage, but could become commercially viable within the next decade. These are based on very low-cost materials and/or processes and include technologies such as dye-sensitized solar cells, organic solar cells and low-cost (printed) versions of existing inorganic thin-film technologies.

Electricity generation by *dye-sensitized solar cells* (DSSCs) is based on light absorption in dye molecules (the “sensitizers”) attached to the very large surface area of a nanoporous oxide semiconductor electrode (e.g.,  $\text{TiO}_2$ ), followed by injection of excited electrons from the dye into the oxide. The dye/oxide interface thus serves as the separator of negative and positive

charges, like the p-n junction in other devices. The injected electrons are then replenished by electrons supplied through a liquid electrolyte which penetrates the pores and which provides the electrical path from the counter electrode [37]. State-of-the-art DSSCs have achieved a top conversion efficiency of 10.9% [12]. Despite the gradual improvements since its discovery in 1991 [38], long-term stability, in combination with a reasonable efficiency (depending on the application) is a key issue in commercializing these PV cells.

*Organic PV* (OPV) cells use stacked solid organic semiconductors, either polymers or small organic molecules. A typical structure of a small molecule OPV cell consists of a stack of p-type and n-type organic semiconductors forming a planar heterojunction. The short-lived nature of the excited states (excitons) formed upon light absorption limits the thickness of the semiconductor layers that can be used and therefore the efficiency of such devices. Note that excitons need to move to the interface where positive and negative charges can be separated before they de-excite. If the travel distance is short, the “active” thickness of the material is small and not all light can be absorbed within that thickness. The efficiency that can be achieved with single-junction OPV cells is about 5% [39]. To decouple exciton transport distances from optical thickness (light absorption), so-called bulk-heterojunction devices have been developed. In these devices, the absorption layer is made of a nanoscale mixture of p- and n-type materials (respectively polymers such as P3HT and fullerenes) to allow the excitons to reach the interface within their lifetime, while also enabling a sufficient macroscopic layer thickness. This bulk-heterojunction structure plays a key role in improving the efficiency, to a record value of 8.3% [12]. The developments in cost and processing [40, 41] of materials have caused OPV research to advance further. Also here the main development challenge is the achievement of a sufficiently high stability in combination with a reasonable efficiency.

*Novel technologies* are potentially disruptive (high-risk, high-potential) approaches based on new materials, devices, and conversion concepts. Generally, their practically achievable conversion efficiencies and cost structure are still unclear. Examples of

these approaches include intermediate band semiconductors, hot carrier devices, spectrum converters, plasmonic solar cells, and various applications of quantum dots. Whereas the emerging technologies described in the previous section primarily aim at very low cost, while achieving a sufficiently high efficiency and stability, many (not all) novel technologies aim at reaching very high efficiencies, by making better use of the entire solar spectrum from infrared to ultraviolet. Generally, their *practically* achievable conversion efficiencies and cost structure are still unclear.

### Photovoltaic System

A *photovoltaic system* is composed of the PV module, as well as the balance of system, which includes storage, system utilization, and the energy network. The system must be reliable, cost-effective, attractive, and match with the electric grid in the future. Detailed information can be found in various studies and roadmaps from organizations around the world (U.S. Photovoltaic Industry Roadmap [42]; Navigant Consulting Inc. [43]; EU PV European Photovoltaic Technology Platform [44]; Kroposki [45]; NEDO [46]).

At the component level, a major objective of balance-of-system (BOS) development is to extend the lifetime of BOS components for grid-connected applications to that of the modules, typically 20–30 years, in addition to further reducing the cost of components and of installation. The highest priority is given to developing inverters, storage devices, and new designs for specific applications such as building-integrated PV. For systems installed in isolated, off-grid areas, component lifetime should be increased to around 10 years, and components for these systems need to be designed so that they require little or no maintenance. Storage devices are necessary for off-grid PV systems and will require innovative approaches to the short-term storage of small amounts of electricity (1–10 kWh); in addition, approaches are needed for integrating the storage component into the module, thus providing a single streamlined product that is easy to use in off-grid and remote applications. Moreover, devices for storing large amounts of electricity (over 1 MWh) will be adapted

to large PV systems in the new energy network. As new module technologies emerge in the future, some of the ideas relating to BOS may need to be revised. Furthermore, the quality of the system needs to be assured and adequately maintained according to defined standards, guidelines, and procedures. To assure system quality, assessing performance is important, including online analysis (e.g., early fault detection) and off-line analysis of PV systems. The knowledge gathered can help to validate software for predicting the energy yield of future module and system technology designs.

To increasingly penetrate the energy network, PV systems must use technology that is compatible with the electric grid and energy supply and demand. System designs and operation technologies must also be developed in response to demand patterns by developing technology to forecast power generation volume and to optimize the storage function. Moreover, inverters must improve the quality of grid electricity by controlling reactive power or filtering harmonics with communication in a new energy network such as the Smart Grid.

### Research and Development in Photovoltaics

With the oil crisis of the 1970s, many countries in the world started solar energy research and development (R&D) programs. A vast amount of the research efforts was and is still concentrated on improving the solar cell and module efficiencies, but research efforts in photovoltaic system components, reliability and manufacturing technologies is of equal importance to drive down costs. A new field of research is emerging, which is not only crucial to photovoltaic generated electricity but all fluctuating renewable energy sources—the integration into existing infrastructures or the design of new smart grid structures. In this entry, the main research issues are shortly described as well as the research activities in some world regions.

### Research Issues

PV modules are the basic building blocks of flat-plate PV systems. Further technological efforts should lead to cost reduction, performance enhancement, and an

improved environmental profile. It is useful to distinguish between technology categories that require specific R&D approaches.

1. Wafer-based crystalline silicon.
2. Existing thin-film technologies.
3. Emerging and novel technologies (including “boosters” to technologies in 1 and 2).

For these three technology categories, the following paragraphs list the R&D topics of highest importance. More details can be found in the different PV roadmaps, for example, Strategic Research Agenda for Photovoltaic Solar Energy Technology [44], U.S. Photovoltaic Industry Roadmap [42], and Japanese PV Roadmap [46, 47].

- *Efficiency, energy yield, stability, and lifetime*

Research often aims at optimization rather than maximization of these parameters, which means that additional costs and gains are critically compared. Since research is primarily aimed at reducing the cost of electricity generation it is important not to focus on initial costs (€/Wp) only, but also on lifecycle gains, that is, actual energy yield (kWh/Wp over the economic or technical lifetime).

- *High productivity manufacturing, including in-process monitoring and control*

Throughput and yield are important parameters in low-cost manufacturing and essential to achieve the cost targets. In-process monitoring and control are crucial tools to increase product quality and yield. Dedicated developments are needed to bring PV manufacturing to maturity.

- *Environmental sustainability*

The energy and materials requirements in manufacturing as well as the possibilities for recycling are important parameters in the overall environmental quality of the product. Further shortening of the energy pay-back time, design for recycling and, ideally, avoiding of the use of critical materials are the most important issues to be addressed here.

- *Applicability*

As discussed in more detail in the paragraphs on BOS and systems, standardization and harmonization are important to bring down the costs of PV. Some of the related

aspects have to be addressed on a module level. In addition, improved ease of installation is partially related to module features. Finally, aesthetic quality of modules (and systems) is an important aspect for large-scale use in the built environment.

4. Concentrator systems are a separate category, because the R&D issues are fundamentally different compared to flat-plate technologies.

It is noted, however, that some of the concepts discussed under “Emerging and novel technologies” might in the end be especially suited for use in concentrator systems.

Concentrated Photovoltaics (CPV) offers a variety of technical solutions and these solutions are given on system level. The research issues can be divided into the following activities:

- Concentrator solar cell manufacturing
- Optical system
- Module assembly and fabrication method of concentrator modules and systems
- System aspects – tracking, inverter, and installation issues

However, it has to be stated once more clearly: CPV is a system approach! Only if all the interconnections between the components are considered the whole system is optimized. This means that the optimized component is not necessarily the best choice for the optimum CPV system. This requires strong interactions between different research groups.

5. Balance-of-System (BoS) components and systems.

It is at the system level that the user meets PV technology and their interest is in a reliable, cost-effective and attractive solution to their energy supply needs. This research agenda concentrates on topics that will achieve one or more of the following:

- Reduce costs at the component and/or system level
- Increase the overall performance of the system, including aspects of increased and harmonized component lifetimes, reduction of performance losses and maintenance of performance levels throughout system life
- Improve the functionality of and the services provided by the system, so adding value to the electricity produced

6. Standards, quality assurance, safety, and environmental aspects.

National and especially local authorities and utilities require that PV systems meet agreed standards (like building standards, including e.g., fire safety requirements and electrical safety standards). In a number of cases, either existing standards or differences in local standards (inverter requirements/settings) or the lack of standards (e.g., PV modules/PV elements not being certified as a building element because of the lack of an appropriate standard) hinder the development of the PV market. Standards and/or guidelines are required for the whole value chain. The development of new and adapted standards and guidelines implies in many cases that dedicated research and development is required.

Quality assurance is an important tool assuring the effective functioning of both individual components in a PV system and of the PV system as a whole. Standards and guidelines are an important basis for quality assurance. In-line production control procedures and guidelines also need to be developed. At the system level, monitoring techniques need to be developed for early fault detection.

Recycling is an important building block to ensure a sustainable PV industry. So far most attention has been paid to recycling of crystalline silicon solar modules. Methods for recycling of thin-film modules and BoS components (in the case where no recycling procedures exist) need to be addressed in the future. LCA studies have become an important tool to evaluate the environmental profile of the various renewable energy sources. In order to assure the position of PV with respect to other sources, reliable LCA data are required. From these data properties like the CO<sub>2</sub> emission per kWh of electricity produced and the energy payback time can be calculated. In addition, the results of LCA data can be used in the design phase of new processes and equipment for cell and module production lines.

### Research Activities

**China** In the National Outlines for Medium and Long-term Planning for Scientific and Technological Development (2006–2020) [48], solar energy is listed as a priority theme: *New and renewable energy*

*technologies*: to develop low-cost, large-scale renewable energy development and utilization technologies, large-scale wind power generation equipment; to develop technology of Photovoltaic cells with high cost-effect ratio and its utilization; to develop solar power generation technology and study integration of solar powered buildings; to develop technologies of fuel cells, hydropower, biomass energy, hydrogen energy, geothermal energy, ocean energy, biogas, etc.

Also the National Medium-and-Long Term Renewable Energy Development Plan [49] has listed solar Photovoltaic power generation as an important developing point. Within the National Basic Research Programme of China, the so-called 973 Programme, there is an additional topic on “Basic research of mass hydrogen production using solar energy.” In January 2010, China’s Ministry of Science and Technology announced the approval of the first two *PV State Key Laboratories* to foster and accelerate the development of PV technologies.

**European Union** In addition to the 27 national programs for market implementation, research and development, the European Union has been funding research (DG RTD) and demonstration projects (DG ENER, formerly DG TREN) with the Research Framework Programmes since 1980. Compared to the combined national budgets, the EU budget is rather small, but it plays an important role in creating a European Photovoltaic Research Area. This is of particular interest and importance, as research for Photovoltaics in a number of Member States is closely linked to EU funds. A large number of research institutions from small University groups to large research centers, covering everything from basic material research to industry process optimization, are involved and contribute to the progress of Photovoltaics.

The European Commission’s Research and Development activities are organized in multi-annual Framework Programmes (FP). In addition to the technology-oriented research projects, there are Marie Curie Fellow-ships and the “Intelligent Energy - Europe” (EIE) Programme. The current 7th EC Framework Programme for Research, Technological Development has a duration of 7 years and runs from 2007 to 2013. The Commission expects the following impacts from the research activities: *Through*

technological improvements and economies of scale, the cost of grid-connected PV electricity in Europe is expected to be lowered to a figure in the range of 0.10–0.25 €/kWh by 2020. Research and development should lead to reduced material consumption, higher efficiencies and improved manufacturing processes, based on environmentally sound processes and cycles.

In 2007, the European Commission initiated the European Strategic Energy Technology Plan (SET-PLAN) [50]. The aim of the SET-Plan is to focus, strengthen, and give coherence to the overall effort in Europe with the objective of accelerating innovation in cutting edge European low carbon technologies. In doing so, it will facilitate the achievement of the 2020 targets and the 2050 vision of the Energy Policy for Europe. The Communication on the SET-Plan states:

- ▶ Europe needs to act now, together, to deliver sustainable, secure and competitive energy. The inter-related challenges of climate change, security of energy supply and competitiveness are multifaceted and require a coordinated response. We are piecing together a far-reaching jigsaw of policies and measures: binding targets for 2020 to reduce greenhouse gas emissions by 20% and ensure 20% of renewable energy sources in the EU energy mix; a plan to reduce EU global primary energy use by 20% by 2020; carbon pricing through the Emissions Trading Scheme and energy taxation; a competitive Internal Energy Market; an international energy policy. And now, we need a dedicated policy to accelerate the development and deployment of cost-effective low carbon technologies.

Within the SET-Plan, Photovoltaics was identified as one of the key technologies and the SET-Plan calls for six different European industry initiatives, one of them being solar. The *Solar Europe Industry Initiative* will focus on large-scale demonstration for Photovoltaics and concentrated solar power and is launched in June 2010.

**India** In 2008, Prime Minister Manmohan Singh announced India's first National Action Plan on Climate Change. To cope with the challenges of Climate Change, India identified eight National Missions aimed to develop and use new technologies. The use of solar energy with Photovoltaics and Concentrating Solar

Power (CSP) is described in the National Solar Mission (NSM). The objective of the National Solar Mission is to establish India as a global leader in solar energy, by creating the policy conditions for its diffusion across the country as quickly as possible [51]. The actions for Photovoltaics in the National Solar Mission call for R&D collaboration, technology transfer, and capacity building.

R&D projects are supported by the Ministry of Non-Conventional Energy Sources (MNRE) for more than three decades. The range of topics includes the development of poly silicon and other materials, development of device fabrication processes and improvements in crystalline silicon solar cell/module technology, development of thin-film solar cell technology (based on amorphous silicon films, cadmium telluride (CdTe) films and copper indium diselenide (CIS) thin films, organic, dye sensitized, and carbon nanotubes). MNRE is also supporting development of photovoltaic systems and components used in manufacture of such systems. For the 11<sup>th</sup> plan period (2008 – 2012), the Ministry has identified so-called *Thrust Areas of R&D in Solar Photovoltaic Technology* with the aim to reduce module costs to INR 120/Wp (€ 2.00) [52].

The identified key areas of R&D and technology development are focused on the development of:

1. Poly silicon and other materials
2. Efficient silicon solar cells
3. Thin films materials and solar cell modules
4. Concentrating PV systems
5. PV system design, with the objective of significantly reducing the ratio of capital cost to conversion efficiency

**Japan** In Japan, the Independent Governmental Entity New Energy Development Organisation (NEDO) is responsible for the Research Programme for Renewable Energies. The current program for Photovoltaics in the frame of Energy and Environment Technologies Development Projects has three main pillars [53]:

- New Energy Technology Development
- Introduction and Dissemination of New Energy and Energy Conservation
- International Projects

One of the dominant priorities, besides the future increase in PV production, is obviously the cost reduction of solar cells and PV systems. In addition to these activities, there are programs on future technology (in and outside NEDO) where participation of Japanese institutes or companies occurs by invitation only. For the participation of non-Japanese partners, there are “future development projects” and the NEDO Joint Research Programme, mainly dealing with non-applied research topics.

Within the *New Energy Technology Development Programme* there are projects on Photovoltaic technology specific issues, problems of grid-connected systems, as well as public solicitation.

The *Introduction and Dissemination of New Energy and Energy Conservation Programme* consists of various promotional and awareness campaign projects.

The *International Projects* mainly focus on neighboring Asian developing countries to promote technological development.

**Korea** Korea’s National PV programs have been based on the second 10-year basic plan for New and Renewable Energy (NRE) RD&D established to enhance the level of self-sufficiency in energy supply, to meet the challenging of climate change and to consolidate infrastructure of NRE industry.

The government budget in 2009 for PV R&D was KRW 70.6 billion (€ 48.7 million), a 25% increase from the previous year. The 32 new and 25 continued projects have been initiated under the five R&D subprograms categorized into

- Strategic R&D
- Basic and Innovative R&D
- Core Technologies Development
- Demonstration
- International Joint Research

The R&D budget for the 32 new projects is KRW 31.2 billion (€ 21.5 million). The representative “Strategic R&D” projects funded newly in 2009 are “Development of commercialization technologies of flexible CuInGaSe<sub>2</sub> thin-film solar cells using metal foil substrates” and “Development of large area dye-sensitized solar cells with high reliability.” The second phase R&D support has been continued in the projects initiated in 2008. This includes “Development of mass production

facilities for c-Si solar cells,” “Development of commercialization technologies of large area silicon and CIGS thin-film modules,” and “High efficiency a-Si/c-Si hetero-junction solar cells.”

**Taiwan** In 2009, the Taiwanese National Energy Programme started with a four year budget of NTD 30 billions for 4 years. Solar PV is one of the energy technologies amongst the 8 different ones which are supported via this program.

The Industrial Technology Research Institute (ITRI), a Government-backed research organization, has drawn up an R&D Strategy for Taiwan with the aim to lower module costs to around 1 US\$/Wp (0.77 €/Wp) between 2015 and 2020. The research topics identified range from efficiency increase in the various wafer-based and thin-film solar cells to concentrator concepts and novel devices. Despite the fact that the national R&D budget should be doubled within the next 4 years from NTD 5 billion (€ 125 million) per year to NTD 10 billion (€ 250 million) per year, it is visible that the main focus is on the industry support to increase production capacities and improved manufacturing technologies.

**United States** The research activities for solar photovoltaics are a part of the US Solar Energy Technologies Programme (SETP or Solar Programme), which aims to develop cost-competitive solar energy systems for America. The current multi-annual work-program runs from 2008 to 2012 [54]. More than \$ 170 million (€ 130.8 million) are spent each year for research and development on the two solar electric technologies which are considered to have the greatest potential to reach cost competitiveness by 2015: photovoltaics and concentrating solar power. The program names as the greatest R&D challenges the reduction of costs, improvement of system performance, and the search for new ways to generate and store energy captured from the sun.

The goal of the Solar Technology Research Plan is to help overcoming the challenges and barriers to massive manufacturing, sales, and installation of PV technology. Multiple technologies are being pursued that are at differing stages of maturity. With an effective combination of the talents in industry, university, and national laboratories, the needed cost, performance,

and reliability goals should be achieved. Specific PV R&D efforts toward achieving these goals include:

1. PV Systems and Module Development
2. PV Materials and Cell Technologies
3. Testing and Evaluation
4. Grid/Building Integration

### The Photovoltaic Industry

In 2010, the photovoltaic world market doubled in terms of *production* to 24 GW. The market for installed systems doubled again and values between 16 and 18 GW were reported by various consultancies and institutions. This mainly represents the grid-connected photovoltaic market. To what extent the off-grid and consumer-product markets are included is unclear. The difference of roughly 6–7 GW has therefore to be explained as a combination of unaccounted off-grid installations (approx. 1–200 MW off-grid rural, approx. 1–200 MW communication/signals, approx. 100 MW off-grid commercial), consumer products (ca. 1–200 MW), and cells/modules in stock.

In addition, the fact that some companies report shipment figures whereas others report production figures adds to the uncertainty. The difficult economic conditions contributed to the decreased willingness to report confidential company data. Nevertheless, the figures show a significant growth of the production, as well as an increasing silicon supply situation.

The announced production capacities, based on a survey of more than 350 companies worldwide, increased, even with difficult economic conditions. Despite the fact that a number of players announced a scale back or cancelation of their expansion plans for the time being, the number of new entrants into the field, notably large semiconductor or energy-related companies overcompensated this. At least on paper the expected production capacities are increasing. Only published announcements of the respective companies and no third source info were used. The cut-off date of the info used was August 2011.

It is important to note that production capacities are often announced, taking into account different operation models such as number of shifts, operating hours per year, etc. In addition, the announcements of the increase in production capacity do not always specify when the capacity will be fully ramped up and

operational. This method has of course the setback that a) not all companies announce their capacity increases in advance and b) that in times of financial tightening, the announcements of the scale back of expansion plans are often delayed in order not to upset financial markets. Therefore, the capacity figures just give a trend, but do not represent final numbers.

If all these ambitious plans can be realized by 2015, China will each have about 67% of the worldwide production capacity of 104 GW followed by Taiwan (15%), Europe (9%), and Japan (7%) (Fig. 2) [55, 56].

All these ambitious plans to increase production capacities, at such a rapid pace, depend on the expectations that markets will grow accordingly. This, however, is the biggest uncertainty, as the market estimates for 2011 vary between 17 and 24 GW, with a consensus value in the 20 GW range. In addition, most markets are still dependent on public support in the form of feed-in tariffs, investment subsidies, or tax-breaks.

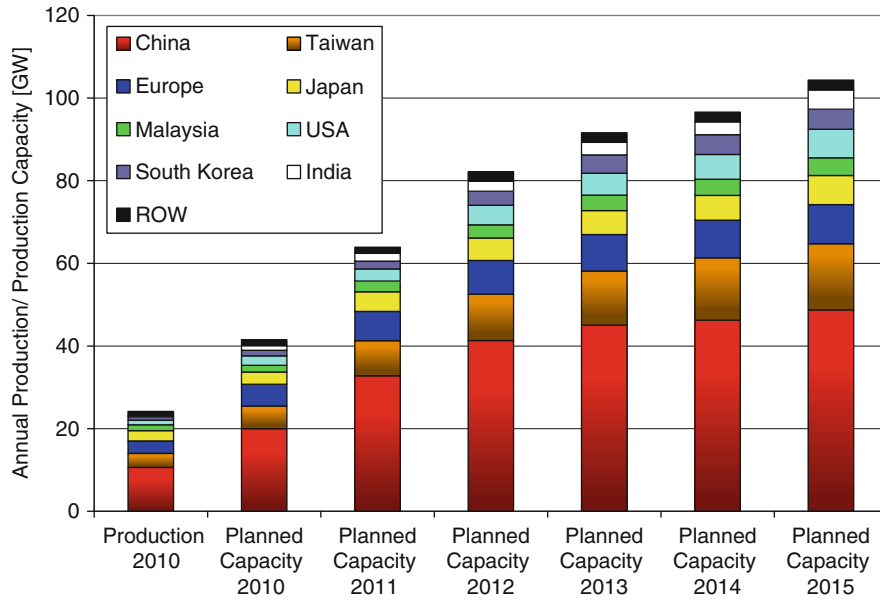
### Technology Mix

Wafer-based silicon solar cells is still the main technology and had around 85% market shares in 2010. Polycrystalline solar cells still dominate the market (45–50%) even if the market shares are decreasing since 2003. Commercial module efficiencies are within a wide range between 12% and 22%, with monocrystalline modules between 14% and 22% and polycrystalline modules between 12% and 18%. The massive capacity increases for both technologies are followed by the necessary capacity expansions for polysilicon raw material.

More than 200 companies are involved in thin-film solar cell activities, ranging from basic R&D activities to major manufacturing activities and over 120 of them have announced the start or increase of production. In 2005, production of Thin-Film solar modules reached for the first time more than 100 MW per annum. The first 100 MW thin-film factories became operational in 2007, followed by the first 1 GW factory in 2010. If all expansion plans are realized in time, thin-film production capacity could be 17 GW, or 21% of the total 82 GW in 2012 and 27 GW, or 26%, in 2015 of a total of 104 GW (Fig. 3).

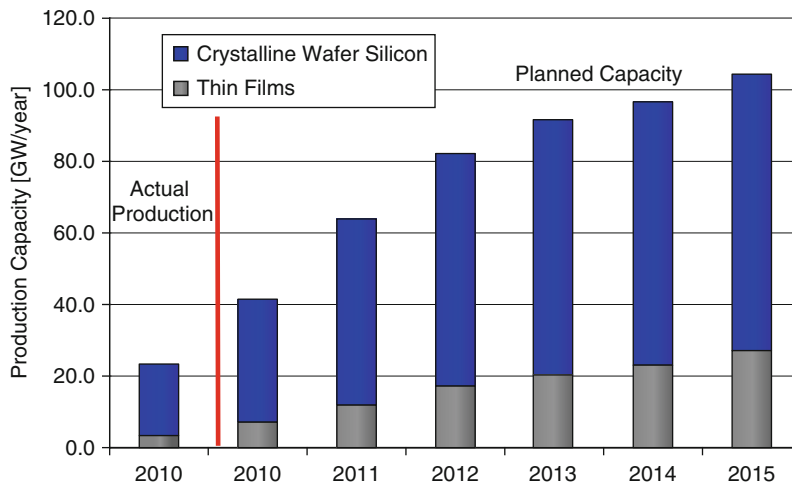
One should bear in mind that only one third of the over 120 companies, with announced production





Photovoltaics, Status of. Figure 2

Worldwide PV Production 2010 with future planned production capacity increases



Photovoltaics, Status of. Figure 3

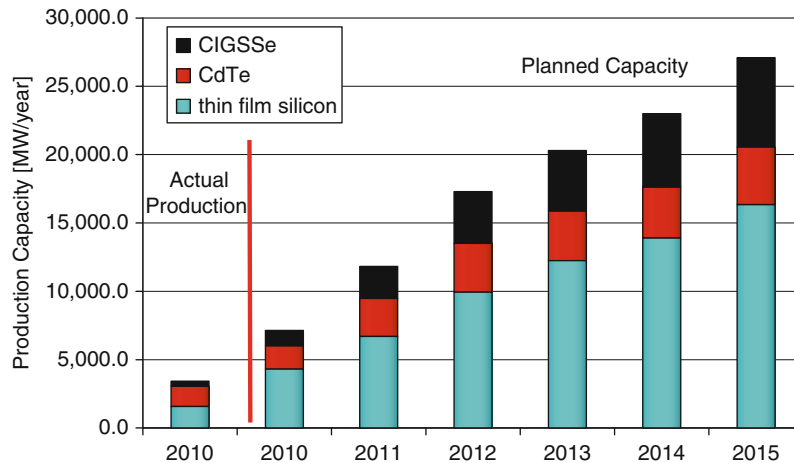
Annual PV production capacities of thin-film and crystalline silicon-based solar modules

plans, have produced thin-film modules of 10 MW or more in 2010 (Fig. 4).

More than 70 companies are silicon-based and use either amorphous silicon or an amorphous/microcrystalline silicon structure. Thirty-six companies announced using Cu (In,Ga) (Se,S)<sub>2</sub> as absorber material for their thin-film solar modules, whereas nine

companies use CdTe and eight companies go for dye and other materials.

Concentrating Photovoltaics (CPV) is an emerging market with approximately 17 MW cumulative installed capacity at the end of 2008. There are two main tracks – either high concentration >300 suns (HCPV) or low to medium concentration with



Photovoltaics, Status of. Figure 4

Annual thin-film PV production capacities of thin-film and crystalline silicon-based solar modules

a concentration factor of 2 to approximately 300. In order to maximize the benefits of CPV, the technology requires high Direct Normal Irradiation (DNI) and these areas have a limited geographical range – the “Sun Belt” of the Earth. The market share of CPV is still small, but an increasing number of companies are focusing on CPV. In 2008, about 10 MW of CPV were produced, market estimates for 2009 are in the 20–30 MW range and for 2010 about 100 MW are expected.

### Price Trends

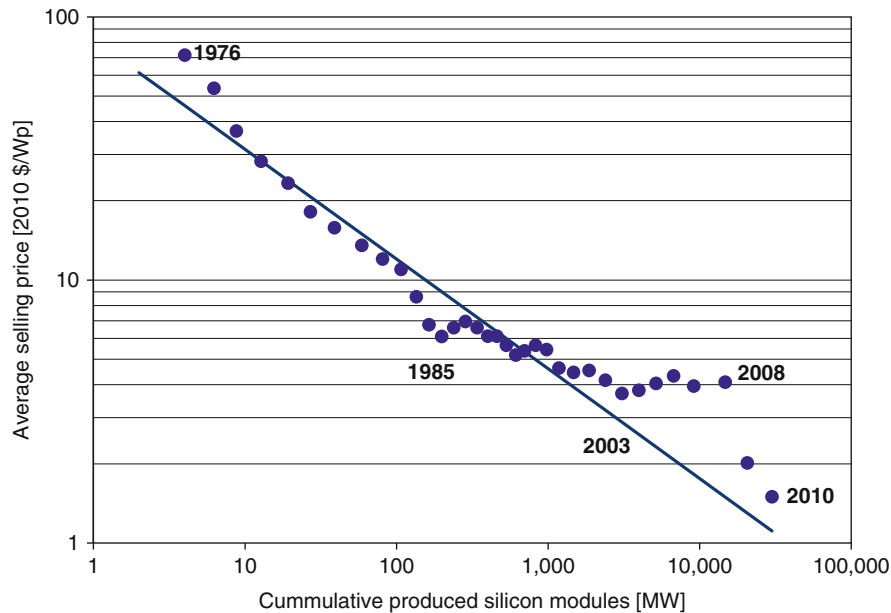
The costs for PV systems and their components have decreased by more than a factor of 10 over the last 30 years. However, if we look at prices, these depend not only on technology development and market size, but also on the global or local competition in the respective market segment. Support schemes, permitting rules, and grid access regulations have a significant influence on prices as well.

The average global PV module factory prices dropped from more than 23 \$/W (2010 \$) in 1980 to about 1.5 \$/W (2010 \$) in 2010 [57]. The majority of studies about learning curve experience in photovoltaics focus on PV modules because they represent the single-largest cost item of a PV system [58]. For the PV modules a range between 11% and 26% is given for the learning experience [59–62] with a median progress ratio of 80%. This leads to a learning rate

(price experience factor) of 20%, which means that the price is reduced by 20% for each doubling of cumulative sales [63, 64]. Figure 5 shows the price developments for crystalline silicon modules over the last 35 years. The huge growth of demand after 2003 led to an increase of prices due to the supply-constrained market, which then changed into a demand-driven market leading to a significant price reduction due to module overcapacities in the market [65].

BOS components are responsible for the second largest technical-related cost part of a PV system. Within the BOS components, the single-largest item is the inverter. While the overall BOS experience curve was between 78% and 81%, or a 19–22% learning rate, quite similar to the module rates learning rates for inverters were just in the range of 10% [67]. A similar trend was found in the USA for cost reduction for labor costs attributed to installed PV systems [68].

The average cost of installed PV systems has also decreased significantly over the past couple of decades and is projected to continue decreasing rapidly as the PV technology and markets mature. However, as already pointed out earlier, the system price decrease varies significantly from region to region and depends strongly on the implemented support schemes and maturity of markets [69]. This study found that the capacity-weighted average costs of PV systems installed in the USA declined from 10.8 \$/W (2010 \$) in 1998 to 7.5 \$/W (2010 \$) in 2009. This decline was attributed primarily to a drop in non-module (BOS) costs.



**Photovoltaics, Status of. Figure 5**

Price experience or learning curve for silicon PV modules. The *straight line* on this log-log plot represents a learning rate of 25% and the actual data follow the supply and demand fluctuations [57, 66]

Figure 6 shows the system price developments in Europe, Japan, and the USA.

Since the second half of 2008, PV system prices have decreased considerably due to the increased competition between PV companies because of huge increases in production capacity and production overcapacities. Already now, electricity production from photovoltaic solar systems has shown that it can be cheaper than peak prices in the electricity exchange.

Average US prices in the first quarter of 2011 were given as: residential systems  $6.41 \pm 2.0$  \$/W ( $4.93 \pm 1.54$  €/W), nonresidential systems  $5.35 \pm 1.65$  \$/W ( $4.12 \pm 1.27$  €/W), and utility scale systems  $3.85 \pm 1.15$  \$/W ( $2.96 \pm 1.04$  €/W) [74]. In the second quarter of 2011, the German average price index, for rooftop systems up to 100 kWp, was given with € 2,422/kWp without tax or half the price of 5 years ago [75]. With such investment costs, the electricity generation costs are already at the level of residential electricity prices in some countries, depending on the actual electricity price, and the local solar radiation level.

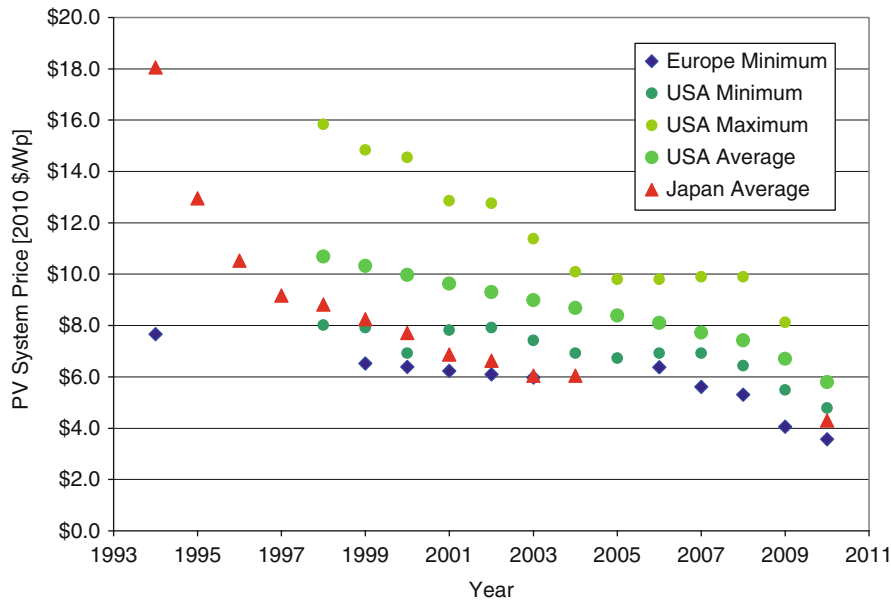
But only if markets and competition continue to grow, prices of the photovoltaic systems will continue to decrease and make electricity from PV systems for

consumers even cheaper than from conventional sources. In order to achieve the price reductions and reach grid parity for electricity generated from photovoltaic systems, public support, especially on regulatory measures, will be necessary for the next decade.

### Solar Cell Production Companies

Worldwide, more than 350 companies produce solar cells. The following subsections give a short description of the 20 largest companies, in terms of actual production/shipments in 2010. More information about solar cell companies and details can be found in various market studies or in the annual *PV Status Report* published by the European Commission's Joint Research Centre [76]. The capacity, production, or shipment data are from the annual reports or financial statements of the respective companies or the cited references.

**Suntech Power Co. Ltd. (PRC)** Suntech Power Co. Ltd. ([www.suntech-power.com](http://www.suntech-power.com)) is located in Wuxi. It was founded in January 2001 by Dr. Zhengrong Shi and went public in December 2005. Suntech specializes in the design, development, manufacturing, and sale of



**Photovoltaics, Status of. Figure 6**

Installed cost of PV systems smaller 100 kW<sub>p</sub> in Europe, Japan, and the USA (Data Sources: [69–73])

photovoltaic cells, modules, and systems. For 2010, Suntech reported shipments of 1507 MW, taking the top rank amongst the solar cell manufacturers. The annual production capacity of Suntech Power was increased to 1.8 GW by the end of 2010, and the company plans to expand its capacity to 2.4 GW in 2011.

**JA Solar Holding Co. Ltd. (PRC)** JingAo Solar Co. Ltd. (<http://www.jasolar.com>) was established in May 2005 by the Hebei Jinglong Industry and Commerce Group Co. Ltd., the Australia Solar Energy Development Pty. Ltd. and Australia PV Science and Engineering Company. Commercial operation started in April 2006 and the company went public on February 7, 2007. According to the company, the production capacity should increase from 1.9 GW at the end of 2010 to 2.5 GW in 2011. For 2010, shipments of 1,460 MW are reported.

**First Solar LLC (USA/Germany/Malaysia)** First Solar LLC (<http://www.firstsolar.com>) is one of the few companies worldwide to produce CdTe-thin-film modules. The company has currently three manufacturing sites in Perrysburg (USA), Frankfurt/Oder (Germany), and in Kulim (Malaysia), which had a combined capacity of 1.5 GW at the end of 2010.

The second Frankfurt/Oder plant, doubling the capacity there to 512 MW, became operational in May 2011 and the expansion in Kulim is on track to increase the production capacity to 2.3 GW at the end of 2011. Further expansions are under way in Meze (AZ), USA, and Dong Nam Industrial Park, Vietnam, to increase the production capacity to 2.9 GW at the end of 2012. The new factory planned in the framework of a joint venture with EDF Nuouvelles in France is currently on hold. In 2010, the company produced 1.4 GW and currently sets the production cost benchmark with 0.75 \$/Wp (0.58 €/Wp) in the first quarter of 2011.

**Sharp Corporation (Japan/Italy)** Sharp ([www.sharp-world.com](http://www.sharp-world.com)) started to develop solar cells in 1959 and commercial production got under way in 1963. Since its products were mounted on “Ume,” Japan’s first commercial-use artificial satellite, in 1974, Sharp has been the only Japanese maker to produce silicon solar cells for use in space. Another milestone was achieved in 1980, with the release of electronic calculators, equipped with single-crystal solar cells.

In 2010, Sharp had a production capacity of 1,070 MWp/year, and shipments of 1.17 GW were reported [79]. Sharp has two solar cell factories in Japan,

Katsuragi, Nara Prefecture (550 MW c-Si and 160 MW a-Si their triple-junction thin-film solar cell), and Osaka (200 MW c-Si and 160 MW a-Si), one together with Enel Green Power and STMicroelectronics in Catania, Italy (initial capacity 160 MW at the end of 2011), six module factories and the Toyama factory to recycle and produce silicon. Three of the module factories are outside Japan, one in Memphis, Tennessee, USA with 100 MW capacity, one in Wrexham, UK, with 500 MW capacity and one in Nakornpathom, Thailand.

**Trina Solar Ltd., PRC (PRC)** Trina Solar (<http://www.trinasolar.com/>) was founded in 1997 and went public in December 2006. The company has integrated product lines, from ingots to wafers and modules. In December 2005, a 30 MW monocrystalline silicon wafer product line went into operation. According to the company, the production capacity was 750 MW for ingots and wafers and 1.2 GW for cells and modules at the end of 2010. For 2011, it is planned to expand the capacities to 1.2 GW for ingots and wafers and to 1.9 GW for cells and modules. For 2010, shipments of 1.06 GW were reported.

In January 2010, the company was selected by the Chinese Ministry of Science and Technology to establish a State Key Laboratory to develop PV technologies within the Changzhou Trina PV Industrial Park. The laboratory is established as a national platform for driving PV technologies in China. Its mandate includes research into PV-related materials, cell and module technologies, and system-level performance. It will also serve as a platform to bring together technical capabilities from the company's strategic partners, including customers and key PV component suppliers, as well as universities and research institutions.

**Yingli Green Energy Holding Company Ltd. (PRC)** Yingli Green Energy (<http://www.yinglisolar.com/>) went public on June 8, 2007. The main operating subsidiary, Baoding Tianwei Yingli New Energy Resources Co. Ltd., is located in the Baoding National High-New Tech Industrial Development Zone. The company deals with the whole set, from solar wafers, cell manufacturing, and module production. According to the company, production capacity reached 1 GW in July 2010. A further expansion project to 1.7 GW is ongoing and

should be operational at the end of 2011. The financial statement for 2010 gave shipments of 1.06 GW.

In January 2009, Yingli acquired Cyber Power Group Limited, a development stage enterprise designed to produce polysilicon. Through its principle operating subsidiary, Fine Silicon, the company started trial production of solar-grade polysilicon in late 2009 and is expected to reach its full production capacity of 3,000 t per year by the end of 2011.

In January 2010, the Ministry of Science and Technology of China approved the application to establish a national-level key laboratory in the field of PV technology development, the State Key Laboratory of PV Technology at Yingli Green Energy's manufacturing base in Baoding.

**Q-Cells AG (Germany/Malaysia)** Q-Cells SE (<http://www.qcells.de>) was founded at the end of 1999 and is based in Thalheim, Sachsen-Anhalt, Germany. Solar cell production started in mid 2001, with a 12 MWp production line. In the 2010 Annual Report, the company stated that the nominal capacity was 1.1 GW by the end of 2010, 500 MW in Germany and 600 MW in Malaysia. In 2010, production was 936 MW, 479 MW in Germany, and 457 MW in Malaysia.

In the first half of the last decade, Q-Cells broadened and diversified its product portfolio by investing in various other companies, or forming joint ventures. Since the first half of 2009, Q-Cells has sold most of these holdings and now has one fully owned solar cell manufacturing subsidiary, Solibro (CIGS) with a 2010 production of 75 MW.

**Motech Solar (Taiwan/PRC)** Motech Solar (<http://www.motech.com.tw>) is a wholly owned subsidiary of Motech Industries Inc., located in the Tainan Science Industrial Park. The company started its mass production of polycrystalline solar cells at the end of 2000, with an annual production capacity of 3.5 MW. The production increased from 3.5 MW in 2001 to 850 MW in 2010. In 2009, Motech started the construction of a factory in China which should reach its nameplate capacity of 500 MW in 2011. Production capacity at the end of 2010 was given as 1.2 GW (860 MW in Taiwan and 340 MW in China).

In 2007, Motech Solar's Research and Development Department was upgraded to Research and

Development Centre (R&D Centre), with the aim not only to improve the present production processes for wafer and cell production, but to develop next generation solar cell technologies.

At the end of 2009, the company announced that it acquired the module manufacturing facilities of GE in Delaware, USA.

**Gintech Energy Corporation (Taiwan)** Gintech (<http://www.gintech.com.tw/>) was established in August 2005 and went public in December 2006. Production at Factory Site A, Hsinchu Science Park, began in 2007 with an initial production capacity of 260 MW and increased to 930 MW at the end of 2010. The company plans to expand capacity to 1.5 GW in 2011. In 2010, the company had a production of 827 MW [3].

**Kyocera Corporation (Japan)** In 1975, Kyocera (<http://global.kyocera.com/prdct/solar/>) began with research on solar cells. The Shiga Yohkaichi Factory was established in 1980 and R&D and manufacturing of solar cells and products started with mass production of multicrystalline silicon solar cells in 1982. In 1993, Kyocera started as the first Japanese company to sell home PV generation systems.

Besides the solar cell manufacturing plants in Japan, Kyocera has module manufacturing plants in China (joint venture with the Tianjin Yiqing Group (10% share) in Tianjin since 2003), Tijuana, Mexico (since 2004), and in Kadan, Czech Republic (since 2005).

In 2010, Kyocera had a production of 650 MW and is also marketing systems that both generate electricity through solar cells and exploit heat from the sun for other purposes, such as heating water. The Sakura Factory, Chiba Prefecture, is involved in everything from R&D and system planning to construction and servicing and the Shiga Factory, Shiga Prefecture, is active in R&D, as well as the manufacturing of solar cells, modules, equipment parts, and devices, which exploit heat. Like solar companies, Kyocera is planning to increase its current capacity of 650 MW in 2010 to 800 MW in 2011 and 1 GW in 2012.

**SunPower Corporation (USA/Philippines/Malaysia)** SunPower (<http://us.sunpowercorp.com/>) was founded in 1988 by Richard Swanson and Robert

Lorenzini to commercialize proprietary high-efficiency silicon solar cell technology. The company went public in November 2005. SunPower designs and manufactures high-performance silicon solar cells based on an interdigitated rear-contact design for commercial use. The initial products, introduced in 1992, were high-concentration solar cells with an efficiency of 26%. SunPower also manufactures a 22% efficient solar cell, called Pegasus, that is designed for non-concentrating applications.

SunPower conducts its main R&D activity in Sunnyvale, California, and has its cell manufacturing plant outside of Manila in the Philippines, with 590 MW capacity (Fab. No 1 and No 2). Fab. No. 3, a joint venture with AU Optronics Corporation (AUO), with a planned capacity of 1.4 GW, is currently under construction in Malaysia. Production in 2010 was reported at 584 MW.

**Canadian Solar Inc. (PRC)** Canadian Solar Inc. was founded in Canada in 2001 and was listed on NASDAQ in November 2006. CSI has established six wholly owned manufacturing subsidiaries in China, manufacturing ingot/wafer, solar cells, and solar modules. According to the company, it had 200 MW of ingot and wafer capacity, 800 MW cell capacity and 1.3 GW module manufacturing capacity in 2010. The company reports that it is on track to expand their solar cell capacity to 1.3 GW and the module manufacturing capacity to 2 GW, including 200 MW in Ontario, Canada, in 2011. For 2010, the company reported production of 522 MW solar cells and sales of 803 MW of modules.

**Hanwah Solar One (PRC/South Korea)** Hanwah Solar One (<http://www.hanwha-solarone.com>) was established in 2004 as Solarfun Power Holdings, by the electricity meter manufacturer, Lingyang Electronics, the largest Chinese manufacturer of electric power meters. In 2010, the Korean company, Hanwha Chemical, acquired 49.99% of the shares and a name change was performed in January 2011. The company produces silicon ingots, wafers, solar cells, and solar modules. The first production line was completed at the end of 2004 and commercial production started in November 2005. The company went public in December 2006 and reported the completion of its production

capacity expansion to 360 MW in the second quarter of 2008.

As of April 30, 2011, the company reported the following capacities: 1 GW PV module production capacity, 700 MW of cell production capacity, 415 MW of ingot production capacity, and 500 MW of wire-sawing capacity. It is planned to expand the module production capacity to 1.5 GW, cell production capacity to 1.3 GW and ingot and wafer production capacity to 1 GW by the end of 2011.

The 2010 annual production was reported with 360 MW ingots, 387 MW wafers, 502 MW solar cells, and 759 modules.

**Neo Solar Power Corporation (Taiwan)** Neo Solar Power (<http://www.neosolarpower.com/>) was founded in 2005 by PowerChip Semiconductor, Taiwan's largest DRAM company, and went public in October 2007. The company manufactures mono- and multicrystalline silicon solar cells and offers their SUPERCELL multicrystalline solar cell brand with 16.8% efficiency. Production capacity of silicon solar cells at the end of 2010 was 820 MW and the expansion to more than 1.3 GW is planned for 2011. In 2010, the company had shipments of about 500 MW.

**Renewable Energy Corporation as (Norway/Singapore)** REC's (<http://www.recgroup.com/>) vision is to become the most cost-efficient solar energy company in the world, with a presence throughout the whole value chain. REC is presently pursuing an aggressive strategy to this end. Through its various group companies, REC is already involved in all major aspects of the PV value chain. The company located in Høvik, Norway, has five business activities, ranging from silicon feedstock to solar system installations.

REC ScanCell is located in Narvik, producing solar cells. From the start-up in 2003, the factory has been continuously expanding. In 2010, production of solar cells was 452 MW, with a capacity at year end of 180 MWp in Norway, and 550 MW in Singapore.

**Solar World AG (Germany/USA)** Since its founding in 1998, Solar World (<http://www.solarworld.de/>) has changed from a solar system and components dealer to a company covering the whole PV value chain, from wafer production to system installations. The company

now has manufacturing operations for silicon wafers, cells and modules in Freiberg, Germany, and Hillsboro (OR), USA. Additional solar module production facilities exist in Camarillo (CA), USA, and since 2008 with a joint venture between Solarworld and SolarPark Engineering Co. Ltd. in Jeonju, South Korea.

For 2010, solar cell production capacities in Germany were reported at 250 MW and 500 MW in the USA. Total cell production in 2010 was 451 MW, with 200 MW coming from Germany and 251 MW from the USA.

In 2003, the Solar World Group was the first company worldwide to implement silicon solar cell recycling. The Solar World subsidiary, Deutsche Solar AG, commissioned a pilot plant for the reprocessing of crystalline cells and modules.

**Sun Earth Solar Power Co. Ltd. (PRC)** Sun Earth Solar Power (<http://www.nbsolar.com/>), or NbSolar, has been part of China's PuTian Group since 2003. The company has four main facilities for silicon production, ingot manufacturing, system integration, and solar system production. According to company information, Sun Earth has imported solar cell and module producing and assembling lines from America and Japan.

In 2007, Sun Earth Solar Power relocated to the Ningbo high-tech zone, with the global headquarters of Sun Earth Solar Power. There the company produces wafers, solar cells, and solar modules. The second phase of production capacity expansion to 350 MW was completed in 2009. Further expansion is planned from 450 MW in 2010, 700 MW in 2011, and 1 GW in 2012. For 2010, shipments of 421 MW were reported [3].

**E-TON Solartech Co. Ltd. (Taiwan)** E-Ton Solartech (<http://www.e-tonsolar.com>) was founded in 2001 by the E-Ton Group; a multinational conglomerate dedicated to producing sustainable technology and energy solutions and was listed on the Taiwan OTC stock exchange in 2006.

At the end of 2010, the production capacity was 560 MW per annum and a capacity increase to 820 MW is foreseen for 2011. Shipments of solar cells were reported at 420 MW for 2010.

**SANYO Electric Company (Japan)** Sanyo (<http://sanyo.com/solar/>) commenced R&D for a-Si solar cells in 1975. In 1980 marked the beginning of Sanyo's a-Si solar cell mass productions for consumer applications. Ten years later in 1990, research on the HIT (Heterojunction with Intrinsic Thin Layer) structure was started. In 1992, Dr. Kuwano (former president of SANYO) installed the first residential PV system at his private home. Amorphous Silicon modules for power use became available from SANYO in 1993 and in 1997 the mass production of HIT solar cells started. In 2010, Sanyo produced 405 MW solar cells [3]. The company announced increasing its 2009 production capacity of 500 MW HIT cells to 650 MW by 2011.

At the end of 2002, Sanyo announced the start of module production outside Japan. The company now has a HIT PV module production at SANYO Energy S.A. de C.V.'s Monterrey, Mexico, and it joined Sharp and Kyocera to set up module manufacturing plants in Europe. In 2005, it opened its module manufacturing plant in Dorog, Hungary.

Sanyo has set a world record for the efficiency of the HIT solar cell, with 23% under laboratory conditions [17]. The HIT structure offers the possibility to produce double-sided solar cells, which has the advantage of collecting scattered light on the rear side of the solar cell and can therefore increase the performance by up to 30%, compared to one-sided HIT modules in the case of vertical installation.

**China Sunergy** China Sunergy was established as CEEG Nanjing PV-Tech Co. (NJPV), a joint venture between the Chinese Electrical Equipment Group in Jiangsu and the Australian Photovoltaic Research Centre in 2004. China Sunergy went public in May 2007. At the end of 2008, the Company had five selective emitter (SE) cell lines, four HP lines, three capable of using multicrystalline and monocrystalline wafers, and one normal P-type line for multicrystalline cells, with a total nameplate capacity of 320 MW. At the end of 2010, the company had a cell capacity of 400 MW and a module capacity of 480 MW. For 2011, a capacity increase to 750 MW cells and 1.2 GW of modules is foreseen. For 2010, a production of 347 MW was reported vThin Film Solar Cell Production Companies.

## Thin Film Solar Cell Production Companies

Worldwide, about 40 companies have produced thin-film modules of 10 MW or more in 2010. The following subsections give a short description of the ten largest companies not yet mentioned in section "Solar Cell Production Companies" (First Solar, Sharp, Solibro as part of Q-Cells), in terms of actual production/shipments in 2010. More information about solar cell companies and details can be found in various market studies or in the annual *PV Status Report* published by the European Commission's Joint Research Centre [76]. The capacity, production, or shipment data are from the annual reports or financial statements of the respective companies or the cited references.

**Trony Solar Holdings Company Ltd** Trony Solar is located in Shenzhen, Guangdong Province, and manufactures thin-film silicon solar cells and modules for BIPV and consumer applications. According to the company, the capacity was 205 MW at the end of 2010. In May 2010, Trony Technology's "1,000 MW Thin-Film Solar Cell Industrial Base" was recognized as one of the "Top 500 Modern Industrial Projects" of Guangdong Province. For 2010, a production of 145 MW is reported [4].

**United Solar Systems** United Solar Systems Corp. is a subsidiary of Energy Conversion Devices, Inc. (ECD). The first 25 MW manufacturing facility of the flexible a-Si triple-junction solar cell is located in Auburn Hills (MI) and was inaugurated in 2002.

According to the company, production capacity was foreseen to expand to 320 MW by 2010 and 720 MW in 2011. In 2008, financing deals were closed which would allow an expansion to 1 GW in 2012 [Ecd 2008]. The current nameplate capacity in Auburn Hills is quoted with 58 MW and in Greenville, Michigan, 120 MW. A joint venture United Solar Ovonic Jinneng Limited was set up with Tianjin Jinneng Investment Company (TJIC) to operate a 30 MW module plant in Tianjin. In April 2011, the company announce to build another module manufacturing plant with an initial capacity of 15 MW in Ontario. Production in 2010 is reported with 120 MW [3].

**Solar Frontier** Solar Frontier is a 100% subsidiary of Showa Shell Sekiyu K.K. In 1986, Showa Shell Sekiyuki



started to import small modules for traffic signals, and started module production in Japan, cooperatively with Siemens (now Solar World). The company developed CIS solar cells and completed the construction of the first factory with 20 MW capacity in October 2006. Commercial production started in FY 2007. In August 2007, the company announced the construction of a second factory with a production capacity of 60 MW to be fully operational in 2009. In July 2008, the company announced they would open a research center “to strengthen research on CIS solar powered cell technology, and to start collaborative research on mass production technology of the solar modules with Ulvac, Inc.” The aim of this project is to start a new plant in 2011 with a capacity of 900 MW. The ramp up started in February 2011. In 2010, the company changed its name to Solar Frontier and production is reported with 74 MW [3].

**Kaneka Solartech** Kaneka has been involved in the development of amorphous solar cells for over 25 years. Initially, this was aimed at the consumer electronics market, but overall R&D, as well as business strategy, changed in 1993 when Kaneka decided to move into the power module market for residential and industrial applications.

Currently Kaneka produces a-Si and amorphous/microcrystalline silicon modules for rooftop application and built-in roofing types for the Japanese, as well as export markets. The built-in roofing types were developed for the Japanese housing market in cooperation with Quarter-House and Kubota and are either shingle type modules or larger roofing elements. In 2006, the company opened a module factory in Olomouc, Czech Republic, where the capacity was increased to 30 MW in 2008. In FY 2010, the total production capacity was expanded to 150 MWp/year. A further expansion to 350 MW in 2011 and to 1 GW in 2015 was announced by the company. In FY 2010, production was 58 MW [3].

**Mitsubishi Heavy Industries** Mitsubishi Heavy Industries (MHI) started their pilot plant production in 2001, because solar energy has attracted increasing attention as an environment-friendly form of energy. In 2010, MHI produced 50 MW of amorphous silicon

solar cells [4] and had a production capacity of 118 MW.

The plasma CVD deposition, used by MHI, allows rapid deposition on large size glass and flexible substrates (roll-to-roll). MHI has stabilized the a-Si single-junction efficiency at 8%, starting with 10% initial efficiency. The degradation process lasts for approximately 3–4 months, before the stabilized efficiency is reached. Long-time outdoor exposure tests performed at JQA showed that the stabilized efficiency does not change and that the lifetime expectancy can be rated at 20–25 years. Mitsubishi is currently working on improving the efficiency to 12% by using a microcrystalline/a-Si structure in the future. Another feature of the Mitsubishi modules is their high voltage. The modules are produced with either 50 V or 100 V and power ratings between 24 and 100 Wp.

**Schott Solar AG (Germany)** Schott Solar AG has been a fully owned subsidiary of Schott AG, Mainz, since 2005, when Schott took over the former joint venture RWE-Schott Solar, except for the Space Solar Cells Division in Heilbronn. Schott Solar’s portfolio comprises crystalline wafers, cells, modules and systems for grid-connected power and stand-alone applications, as well as a wide range of ASI<sup>®</sup> thin-film solar cells and modules. In 2010, the company had a production capacity of 350 MW and an actual production of 320 MW [3].

Development of amorphous silicon solar cells started at MBB in 1980. Phototronics (PST) was founded in 1988. In 1991, one of the world’s first large-area pilot production facilities for amorphous silicon was built. In January 2008, the company started shipments of modules from its new 33 MW manufacturing facility for amorphous silicon thin-film solar modules in Jena, Germany.

**Sunwell Solar Corporation** Sunwell Solar is a subsidiary of CMC Magnetics Corporation, Taiwan’s top compact disk maker, contracted a 45 MW thin-film PV production plant with Oerlikon Solar. The plant started production at the beginning of September 2008. For 2010, a production of 45 MW is reported [4].

**Auria Solar Co** Auria was founded in October 2007 as a joint venture between E-Ton Solar, Lite-On Technology Corp, Hermes-Epitek Corp. and the MiTAC-SYNNEX Group to manufacture thin-film solar cells. The company has chosen Oerlikon as equipment supplier and plans to produce amorphous/micromorph silicon thin-films. The first factory, with a capacity of 60 MW, began pilot production at the end of 2008 and ramped up to full capacity in 2019. In March 2011, the company announced a cooperation agreement with Mitsubishi Heavy Industries (Japan) for their next expansion to 200 MW in 2012. For 2010, a production of 40 MW is reported [4].

**Baoding TianWei SolarFilms Co., Ltd** Baoding TianWei Solar Films was set up in 2008. It is a subsidiary of the Baoding TianWei Group Co., Ltd., a leading company in the China power transformer industry. Phase I of the production was set up with a capacity of 50 MW and the start of commercial operation was in the second half of 2009. The company plans to reach a capacity of 500 MW in 2015. For 2010, a production of 40 MW is reported [4].

**Moser Baer** Moser Baer Photovoltaic Limited (MBPV) and PV Technologies India Limited (PVTIL) are subsidiaries of Moser Baer India Limited. They were launched between 2005 and 2007 with the primary objective of providing reliable solar power as a competitive nonsubsidized source of energy.

At the end of 2010, the production capacity was given by the company with 90 MW crystalline cells, 100 MW crystalline modules, and 50 MW thin-films with expansion plans in place.

### Polysilicon Supply

The rapid growth of the PV industry since 2000 led to the situation where, between 2004 and early 2008, the demand for *polysilicon* outstripped the supply from the semiconductor industry. Prices for purified silicon started to rise sharply in 2007 and in 2008 prices for polysilicon peaked around 500 \$/kg and consequently resulted in higher prices for PV modules. This extreme price hike triggered a massive capacity expansion, not only of established companies, but many new entrants as well. In 2009, more than 90% of total polysilicon,

for the semiconductor and photovoltaic industry, was supplied by seven companies: Hemlock, Wacker Chemie, REC, Tokuyama, MEMC, Mitsubishi, and Sumitomo. However, it is estimated that now about seventy producers are present in the market.

The massive production expansions, as well as the difficult economic situation, led to a price decrease throughout 2009, reaching about 50–55 \$/kg at the end of 2009, with a slight upward tendency throughout 2010 and early 2011.

For 2010, about 140,000 t of solar-grade silicon production were reported, sufficient for around 20 GW, under the assumption of an average materials need of 7 g/Wp [77]. China produced about 45,000 t, or 32%, capable of supplying about 75% of the domestic demand [78]. According to the Semi PV Group Roadmap, the Chinese production capacity rose to 85,000 t of polysilicon in 2010.

In January 2011, the Chinese Ministry of Industry and Information Technology tightened the rules for polysilicon factories. New factories must be able to produce more than 3,000 t of polysilicon a year and meet certain efficiency, environmental, and financing standards. The maximum electricity use is 80 kWh/kg of polysilicon produced a year, and that number will drop to 60 kWh at the end of 2011. Existing plants that consume more than 200 kWh/kg of polysilicon produced at the end of 2011 will be shut down.

Projected silicon production capacities available for solar in 2012 vary between 250,000 t [11] and 410,665 t [79]. The possible solar cell production will in addition depend on the material use per Wp. Material consumption could decrease from the current 7–8 g/Wp down to 5–6 g/Wp, but this might not be achieved by all manufacturers.

**Silicon Production Processes** The high growth rates of the photovoltaic industry and the market dynamics forced the high-purity silicon companies to explore process improvements mainly for two chemical vapor deposition (CVD) approaches – an established production approach known as the Siemens process, and a manufacturing scheme based on fluidized bed (FB) reactors. Improved versions of these two types of processes will very probably be the workhorses of the polysilicon production industry for the near future.

*Siemens process:* In the late 1950s, the Siemens reactor was developed and has been the dominant production route since. About 80% of total polysilicon manufactured worldwide was made with a Siemens-type process in 2009. The Siemens process involves deposition of silicon from a mixture of purified silane or trichlorosilane gas with an excess of hydrogen onto high-purity polysilicon filaments. The silicon growth then occurs inside an insulated reaction chamber or “bell jar,” which contains the gases. The filaments are assembled as electric circuits in series and are heated to the vapor deposition temperature by an external direct current. The silicon filaments are heated to very high temperatures between 1,100°C and 1,175°C at which trichlorosilane with the help of the hydrogen decomposes to elemental silicon and deposits as a thin-layer film onto the filaments. Hydrogen Chloride (HCl) is formed as a by-product.

The most critical process parameter is temperature control. The temperature of the gas and filaments must be high enough for the silicon from the gas to deposit onto the solid surface of the filament, but well below the melting point of 1,414°C that the filament do not start to melt. Second, the deposition rate must be well controlled and not too fast because otherwise the silicon will not deposit in a uniform, polycrystalline manner, making the material unsuitable for semiconductor and solar applications.

*Fluidized bed process:* A number of companies develop polysilicon production processes based on fluidized bed (FB) reactors. The motivation to use the FB approach is the potentially lower energy consumption and a continuous production compared to the Siemens batch process. In this process, tetrahydrosilane or trichlorosilane and hydrogen gases are continuously introduced to the bottom of the FB reactor at moderately elevated temperatures and pressures. At a continuous rate high-purity silicon seeds are inserted from the top and are suspended by the upward flow of gases. At the operating temperatures of 750°C, the silane gas is reduced to elemental silicon and deposits on the surface of the silicon seeds. The growing seed crystals fall to the bottom of the reactor where they are continuously removed.

MEMC Electronic Materials, a silicon wafer manufacturer, has been producing granular silicon from silane feedstock using a fluidized bed approach for

over a decade. Several new facilities will also feature variations of the FB. Several major players in the polysilicon industry, including Wacker Chemie and Hemlock, are developing FB processes, while at the same time continuing to produce silicon using the Siemens process as well.

*Upgraded metallurgical grade (UMG) silicon* was seen as one option to produce cheaper solar-grade silicon with 5- or 6-nines purity, but the support for this technology is waning in an environment where higher-purity methods are cost competitive. A number of companies delayed or suspended their UMG-silicon operations as a result of low prices and lack of demand for UMG material for solar cells.

### Polysilicon Manufacturers

Worldwide more than 100 companies produce or start-up polysilicon production. The following subsections give a short description of the ten largest companies in terms of production capacity in 2010. More information about polysilicon companies and details can be found in various market studies or in the annual *PV Status Report* published by the European Commission's Joint Research Centre [49].

**Hemlock Semiconductor Corporation (USA)** Hemlock Semiconductor Corporation (<http://www.hscpoly.com>) is based in Hemlock, Michigan. The corporation is a joint venture of Dow Corning Corporation (63.25%) and two Japanese firms, Shin-Etsu Handotai Company, Ltd. (24.5%) and Mitsubishi Materials Corporation (12.25%). The company is the leading provider of polycrystalline silicon and other silicon-based products used in the semiconductor and solar industry.

In 2007, the company had an annual production capacity of 10,000 t of polycrystalline silicon and production at the expanded Hemlock site (19,000 t) started in June 2008. A further expansion at the Hemlock site, as well as a new factory in Clarksville, Tennessee, was started in 2008 and brought total production capacity to 36,000 t in 2010. A further expansion to 40,000 t in 2011 and 50,000 t in 2012 is planned [79].

**Wacker Polysilicon (Germany)** Wacker Polysilicon AG (<http://www.wacker.com>) is one of the world's

leading manufacturers of hyper-pure polysilicon for the semiconductor and photovoltaic industry, chlorosilanes, and fumed silica. In 2010, Wacker increased its capacity to over 30,000 t and produced 30,500 t of polysilicon. The next 10,000 t expansion in Nünchritz (Saxony), Germany, started production in 2011. In 2010, the company decided to build a polysilicon plant in Tennessee with 15,000 t capacity. The groundbreaking of the new factory was in April 2011 and the construction should be finished at the end of 2013.

**OCI Company (South Korea)** OCI Company Ltd. (formerly DC Chemical) (<http://www.oci.co.kr/>) is a global chemical company with a product portfolio spanning the fields of inorganic chemicals, petro and coal chemicals, fine chemicals, and renewable energy materials. In 2006, the company started its polysilicon business and successfully completed its 6,500 metric ton P1 plant in December 2007. The 10,500 metric ton P2 expansion was completed in July 2009 and P3 with another 10,000 t brought the total capacity to 27,000 t at the end of 2010. The debottlenecking of P3, foreseen in 2011, should then increase the capacity to 42,000 t at the end of the year. Further capacity expansions P4 (20,000 t by 2012) and P5 (24,000 t by 2013) have already started (P4) or will commence in the second half of this year (P5).

**GCL-Poly Energy Holdings Limited (PRC)** GCL-Poly (<http://www.gcl-poly.com.hk>) was founded in March 2006 and started the construction of their Xuzhou polysilicon plant (Jiangsu Zhongneng Polysilicon Technology Development Co. Ltd.) in July 2006. Phase I has a designated annual production capacity of 1,500 t and the first shipments were made in October 2007. Full capacity was reached in March 2008. At the end of 2010, polysilicon production capacity had reached 21,000 t and further expansions to 46,000 t in 2011 and 65,000 t in 2012 are underway. For 2010, the company reported a production 17,850 t of polysilicon.

In August 2008, a joint venture, Taixing Zhongneng (Far East) Silicon Co. Ltd., started pilot production of trichlorosilane. Phase I will be 20,000 t, to be expanded to 60,000 t in the future.

**MEMC Electronic Materials Inc. (USA)** MEMC Electronic Materials Inc. (<http://www.memc.com/>) has its headquarters in St. Peters, Missouri. It started operations in 1959 and the company's products are semiconductor-grade wafers, granular polysilicon, ultra-high purity silane, trichlorosilane (TCS), silicon tetrafluoride (SiF<sub>4</sub>), and sodium aluminum tetrafluoride (SAF). MEMC's production capacity in 2008 was increased to 8,000 t and 9,000 t in 2009 [79].

**Renewable Energy Corporation as (Norway)** REC's (<http://www.recgroup.com/>) vision is to become the most cost-efficient solar energy company in the world, with a presence throughout the whole value chain. REC is presently pursuing an aggressive strategy to this end. Through its various group companies, REC is already involved in all major aspects of the PV value chain. The company located in Høvik, Norway has five business activities, ranging from silicon feedstock to solar system installations.

In 2005, Renewable Energy Corporation AS ("REC") took over Komatsu's US subsidiary, Advanced Silicon Materials LLC ("ASiMI"), and announced the formation of its silicon division business area, "REC Silicon Division," comprising the operations of REC Advanced Silicon Materials LLC (ASiMI) and REC Solar-Grade Silicon LLC (SGS). Production capacity at the end of 2010 was around 17,000 t [79] and according to the company, 11,460 t electronic grade silicon was produced in 2010.

**LDK Solar Co. Ltd. (PRC)** LDK (<http://www.ldksolar.com/>) was set up by the Liouxin Group, a company which manufactures personal protective equipment, power tools, and elevators. With the formation of LDK Solar, the company is diversifying into solar energy products. LDK Solar went public in May 2007. In 2008, the company announced the completion of the construction and the start of polysilicon production in its 1,000 t polysilicon plant. According to the company, the total capacity was 12,000 t at the end of 2010 which will be increased to 25,000 t in 2011. In 2010, polysilicon production was reported at 5,050 t.

**Tokuyama Corporation (Japan)** Tokuyama (<http://www.tokuyama.co.jp/>) is a chemical company involved in the manufacturing of solar-grade silicon, the base

material for solar cells. The company is one of the world's leading polysilicon manufacturers and produces roughly 16% of the global supply of electronics and solar-grade silicon. According to the company, Tokuyama had an annual production capacity of 5,200 t in 2008 and has expanded this to 9,200 t in 2010. In February 2011, the company broke ground for a new 20,000 ton facility in Malaysia. The first phase with 6,200 t should be finished in 2013.

A verification plant for the vapor to liquid-deposition process (VLD method) of polycrystalline silicon for solar cells has been completed in December 2005. According to the company, steady progress has been made with the verification tests of this process, which allows a more effective manufacturing of polycrystalline silicon for solar cells.

Tokuyama has decided to form a joint venture with Mitsui Chemicals, a leading supplier of silane gas. The reason for this is the increased demand for silane gas, due to the rapid expansion of amorphous/microcrystalline thin-film solar cell manufacturing capacities.

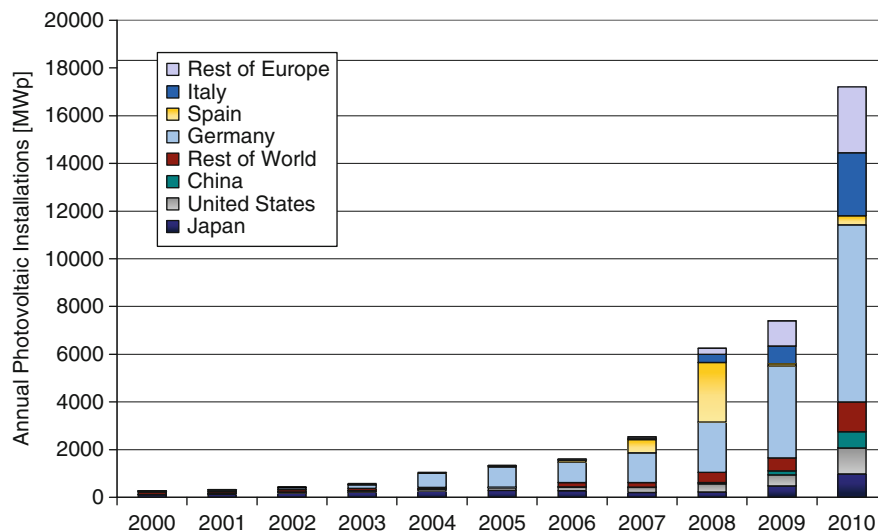
**Kumgang Korea Chemical Company (South Korea)**  
Kumgang Korea Chemical Company (KCC) (<http://www.kccworld.co.kr/english/>) was established by a merger of Kumgang and the Korea Chemical Co. in 2000. In February 2008, KCC announced its investment in the polysilicon industry and began to manufacture

high-purity polysilicon with its own technology at the pilot plant of the Daejuk factory in July of the same year. In February 2010, KCC started to mass-produce polysilicon, with an annual capacity of 6,000 t.

**Mitsubishi Materials Corporation (Japan)**  
Mitsubishi Materials (<http://www.mmc.co.jp>) was created through the merger Mitsubishi Metal and Mitsubishi Mining and Cement in 1990. Polysilicon production is one of the activities in their Electronic Materials and Components business unit. The company has two production sites for polysilicon, one in Japan and one in the USA (Mitsubishi Polycrystalline Silicon America Corporation) and is a shareholder (12.25%) in Hemlock Semiconductor Corporation. With the expansion of the Yokkachi, Mie, Japan, polysilicon plant, by 1,000 t in 2010, total production capacity was increased to 4,300 t.

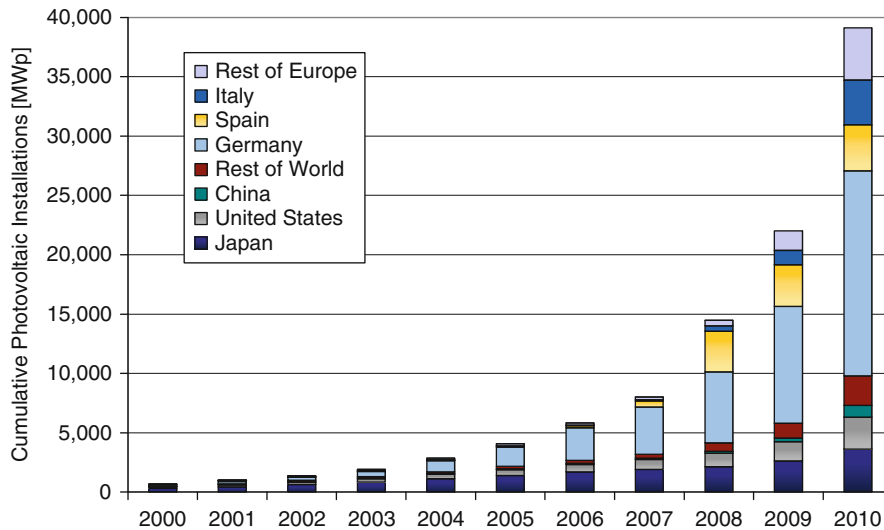
## The Photovoltaic Market

In 2010, the worldwide photovoltaic market more than doubled, driven by major increases in Europe. For 2010, the market volume of newly installed solar photovoltaic electricity systems varies between 17 and 19 GW, depending on the reporting consultancies (Fig. 7). This represents mostly the grid-connected photovoltaic market. To what extent the off-grid and consumer



Photovoltaics, Status of. Figure 7

Annual photovoltaic installations from 2000 to 2010 (Data Source: EPIA [80], Euroobserver [81] and own analysis)



**Photovoltaics, Status of. Figure 8**

Cumulative photovoltaic installations from 2000 to 2010 (Data Source: EPIA [80], Euroserver [81] and own analysis)

product markets are included is not clear, but it is believed that a substantial part of these markets are not accounted for, as it is very difficult to track them. A conservative estimate is that they account for approximately 400–800 MW (approx. 1–200 MW off-grid rural, approx. 1–200 MW communication/signals, approx. 100 MW off-grid commercial and approx. 1–200 MW consumer products).

With a cumulative installed capacity of over 29 GW, the European Union is leading in PV installations with a little more than 70% of the total worldwide 39 GW of solar photovoltaic electricity generation capacity at the end of 2010 (Fig. 8).

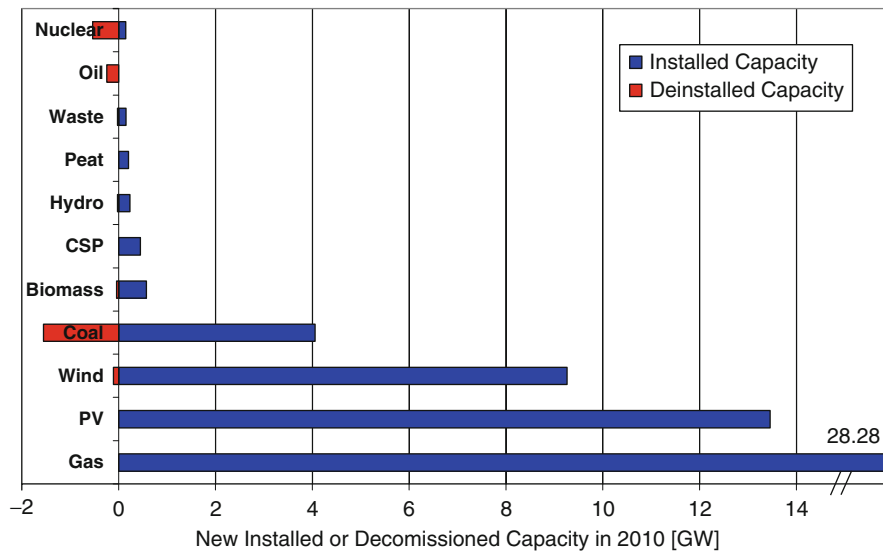
### European Union and Turkey

Market conditions for photovoltaics differ substantially from country to country. This is due to different energy policies and public support programs for renewable energies and especially photovoltaics, as well as the varying grades of liberalization of domestic electricity markets. Within one decade, the solar photovoltaic electricity generation capacity has increased 160 times from 185 MW in 2000 to 29.5 GW in 2010 [80–83].

A total of about 58.8 GW of new power capacity was constructed in the EU last year and 2.5 GW were decommissioned, resulting in 56.3 GW of new net

capacity (Fig. 9) [81, 84]. Gas-fired power stations accounted for 28.3 GW, or 48% of the newly installed capacity. However, in the recent years about 30 GW of additional gas-fired power station projects were suspended or canceled according to Platts [85]. Solar photovoltaic systems moved to the second place with 13.5 GW (23%), followed by 9.4 GW (16%) wind power; 4.1 GW (7%) coal-fired power stations; 570 MW (>1%) biomass; 450 MW (>1%) CSP, 210 MW (>1%) hydro, 230 MW (>1%) peat, and 150 MW (>1%) waste. The net installation capacity for oil-fired and nuclear power plants was negative, with a decrease of 245 MW and 390 MW respectively. The renewable share of new power installations was 40% in 2010.

**Belgium** Belgium showed another strong market performance year in 2010, with new photovoltaic system installations of 420 MW bringing the cumulative installed capacity to 790 MW. However, most of the installations were done in Flanders, where since January 1, 2006 Green Certificates exist with 0.45 €/kWh for 20 years. In Brussels and Wallonia, the Green Certificates have a guaranteed minimum price between 0.15 and 0.65 €/kWh, depending on the size of the systems and region (Brussels 10 years, Wallonia 15 years).



**Photovoltaics, Status of. Figure 9**

New installed or decommissioned electricity generation capacity in Europe in 2010 [81, 84]

**Czech Republic** In the Czech Republic, photovoltaic systems with about 1.5 GW capacity, were installed in 2010, bringing the cumulative nominal capacity to 1.95 GW exceeding their own target of 1.65 GW set in the National Renewable Action Plan for 2020. The Law on the Promotion of Production of Electricity from Renewable Energy Sources went into effect on August 1, 2005 and guarantees a feed-in tariff for 20 years. The annual prices are set by the Energy Regulator. The electricity producers can choose from two support schemes, either fixed feed-in tariffs or market price + Green Bonus. The 2010 feed-in rate in the Czech Republic was CZK 12.25 per kilowatt hour (0.48 €/kWh).

On February 3, 2010, the Czech transmission system operator, EPS, requested all main distribution system operators (EZ, E-ON, PRE) to stop permitting new renewable energy power plants, due to a virtual risk of instability of the electricity grid caused by intermittent renewable sources, especially photovoltaic and wind. Distribution System Operators (DSO) met the requirement on February 16, 2011. The moratorium still exists and SRES (Czech Association of Regulated Energy Companies) announced that the moratorium will continue until at least September.

A number of legislative changes took place in the second half of the year, which resulted in a lower feed-

in tariff for systems larger than 30 kW (5.5 CZK/kWh or 0.216 €/kWh), the phase-out of ground-mounted PV systems from March 1, 2011 onward and the introduction of a retroactive tax on benefits generated by PV installations.

**France** In 2010, 720 MW of PV systems were connected to the grid in France, including about 100 MW which were already installed in 2009. This led to an increase of the cumulative installed capacity to 1.05 GW. However, this positive development came to a sudden stop when the French Prime Minister declared a three-month moratorium on new PV installations above 3 kW and a suspension of projects waiting for grid connection in December 2010.

This rapid growth led to a revision of the feed-in scheme in February 2011, setting a cap of 500 MW for 2011 and 800 MW for 2012 [86]. The new tariff levels only apply to rooftop systems up to 100 kW in size. In addition, those installations are divided into three different categories: residential; education or health; and other buildings with different feed-in tariffs, depending on the size and type of installation. The tariffs for these installations range between 0.2883 €/kWh and 0.46 €/kWh. All other installations up to 12 MW are just eligible for a tariff of 0.12 €/kWh.

**Germany** Germany had the biggest market with 7.4 GW [82]. The German market growth is directly correlated to the introduction of the Renewable Energy Sources Act or “*Erneuerbare Energien Gesetz*” (EEG) in 2000 [87]. This Law introduced a guaranteed feed-in tariff for electricity generated from solar photovoltaic systems for 20 years and already had a fixed built in annual decrease, which was adjusted over time to reflect the rapid growth of the market and the corresponding price reductions. Due to the fact that until 2008 only estimates of the installed capacity existed, a plant registrar was introduced from January 1, 2009 on.

The German market showed two installation peaks during 2010. The first one was in June, when more than 2.1 GW were connected to the grid prior to the 13% feed-in cut which took effect on July 1, 2010. The second peak was in December with almost 1.2 GW just before the scheduled tariff reduction of another 13% on January 1, 2011. Compared to 2009, the feed-in tariff has been reduced by 33–36% depending on the system size and classification. In June 2011, the Bundesnetzagentur (German Federal Network Agency) announced the results of the PV system installation projection required under the Renewable Energy Sources Act (EEG) in order to determine the degression rates for the feed-in tariffs [88]. According to the Agency approximately 700 MW of PV systems were commissioned between March and May 2011 resulting in a projected annual growth of 2.8 GW, which is below the 3.5 GW threshold set for an additional reduction of the tariffs starting July 2011.

**Greece** Greece introduced a new feed-in tariff scheme on January 15, 2009. The tariffs remained unchanged until August 2010 and are guaranteed for 20 years. However, if a grid-connection agreement was signed before that date, the unchanged FIT was applied if the system is finalized within the next 18 months. For small rooftop PV systems, an additional program was introduced in Greece on June 4, 2009. This program covers rooftop PV systems up to 10 kWp (both for residential users and small companies). In 2011, the tariffs decreased by 6.8–8.5%, depending on the size and location of the installation. In 2010, about 150 MW of new installations were carried out, bringing the total capacity to about 205 MW.

**Italy** Italy again took the second place, with respect to new installations and added a capacity of about 2.5 GW, bringing cumulative installed capacity to 3.7 GW at the end of 2010 [83]. In the middle of August 2011, the total connected PV capacity has surpassed 9.4 GW [89]. The *Quarto Conto Energia* (Fourth Energy Bill) was approved by the Italian Council of Ministers on May 5, 2011 [Gaz 2011]. The Bill introduced monthly reductions of the tariffs, starting from June 2011 until January 2012 and then another one in July 2012. In addition, the new Bill limits the feed-in tariffs for new systems up until the end of 2016, or until a cap of 23 GW is reached. In addition, separate caps for large systems are set for the second half of 2011 (1.35 GW) and 2012 (1.75 GW).

**Spain** Spain is second regarding the total cumulative installed capacity with 3.9 GW. Most of this capacity was installed in 2008 when the country was the biggest market, with close to 2.7 GW in 2008 [80]. This was more than twice the expected capacity and was due to an exceptional race to install systems before the Spanish Government introduced a cap of 500 MW on the yearly installations in the autumn of 2008. A revised Decree (Royal Decree 1758/2008) set considerably lower feed-in tariffs for new systems and limited the annual market to 500 MW, with the provision that two thirds are rooftop mounted and no longer free-field systems. These changes resulted in a new installed capacity of about 100 MW and about 380 MW in 2010.

In 2010, the Spanish Government passed the Royal Decrees 1565/10 [90] and RD-L 14/10 [91]. The first one limits the validity of the feed-in tariffs to 28 years, while the latter reduces the tariffs by 10–30% for existing projects until 2014. Both Bills are “retroactive” and the Spanish Solar Industry Association (ASIF) [92] has already announced taking legal actions against them.

**United Kingdom** The United Kingdom introduced of a new feed-in tariff scheme in 2010, which led to the installation of approximately 55 MW, bringing the cumulative installed capacity to about 85 MW. However, in March 2011, the UK Government proposed significant reductions of the tariffs, especially for systems larger than 50 kW.



**Other European Countries and Turkey** Despite high solar radiation, solar photovoltaic system installation in *Portugal* has only grown very slowly and reached a cumulative capacity of 130 MW at the end of 2010.

The market in *Slovakia* showed an unexpected growth from less than 1 MW installed at the end of 2009 to about 144 MW at the end of 2010. In December 2010, the Slovak Parliament adopted an Amendment to the Renewable Energy Sources (RES) Promotion Act, decreasing the feed-in tariffs and from February 1, 2011 on only solar rooftop facilities or solar facilities on the exterior wall of buildings, with capacity not exceeding 100 kW, are eligible for the feed-in tariff. As a result, larger new solar projects in *Slovakia* are on hold.

In *Turkey* in March 2010, the Energy Ministry unveiled its 2010–2014 Strategic Energy Plan. One of Government's priorities is to increase the ratio of renewable energy resources to 30% of total energy generation by 2023. At the beginning of 2011, the Turkish Parliament passed a Renewable Energy Legislation which defines new guidelines for feed-in tariffs. The feed-in tariff is 0.133 \$/kWh (0.10 €/kWh) for owners commissioning a PV system before the end of 2015. If components “*Made in Turkey*” are used, the tariff will increase by up to \$0.067 (€ 0.052), depending on the material mix. Feed-in tariffs apply to all types of PV installations, but large PV power plants will receive subsidies up to a maximum size of 600 MWp.

### Asia and Pacific Region

The Asia and Pacific Region shows an increasing trend in photovoltaic electricity system installations. There are a number of reasons for this development, ranging from declining system prices, heightened awareness, favorable policies, and the sustained use of solar power for rural electrification projects. Countries such as Australia, China, India, Indonesia, Japan, Malaysia, South Korea, Taiwan, Thailand, The Philippines and Vietnam show a very positive upward trend, thanks to increasing governmental commitment toward the promotion of solar energy and the creation of sustainable cities.

The introduction or expansion of feed-in-tariffs is expected to be an additional big stimulant for on-grid solar PV system installations for both distributed and

centralized solar power plants in countries such as Australia, Japan, Malaysia, Thailand, Taiwan, and South Korea.

The Asian Development Bank (ADB) launched an Asian Solar Energy Initiative (ASEI) in 2010, which should lead to the installation of 3 GW of solar power by 2012 [93]. In their report, ADB states: *Overall, ASEI aims to create a virtuous cycle of solar energy investments in the region, toward achieving grid parity, so that ADB developing member countries optimally benefit from this clean, inexhaustible energy resource.*

**Australia** In 2010, 383 MW of new solar photovoltaic electricity systems were installed in Australia, bringing the cumulative installed capacity of grid-connected PV systems to 571 MW [94]. The 2010 market was dominated by the increase of grid-connected distributed systems, which increased from 67 MW in 2009 to 378 MW in 2010. The newly installed PV electricity generation capacity in Australia accounted for 20% of the new electricity generation capacity in 2010.

Most installations took advantage of the incentives under the Australian Government's Solar Homes and Communities Plan (SHCP), Renewable Energy Target (RET) mechanisms and feed-in tariffs in some States or Territories. At the beginning of 2010, eight out of the eleven Australian Federal States and Territories had introduced some kind of feed-in tariff scheme for systems smaller than 10 kWp. All of these schemes have built-in caps which were partly reached that year so that in 2011 only five State schemes are still available for new installations and additional changes are expected in the course of this year.

**India** For 2010, market estimates for solar PV systems vary between 50 and 100 MW, but most of these capacities are for off-grid applications. The Indian National Solar Mission was launched in January 2010, and it was hoped that it would give impetus to the grid-connected market, but only a few MW were actually installed in 2010. The majority of the projects announced will come online from 2011 onward.

The National Solar Mission aims to make India a global leader in solar energy and envisages an installed solar generation capacity of 20 GW by 2020, 100 GW by 2030, and 200 GW by 2050. The short-term outlook up

until 2013 was improved as well when the original 50 MW grid-connected PV system target in 2012 was changed to 1,000 MW for 2013.

**Japan** In 2010, the Japanese market experienced a high growth, doubling its volume to 990 MW, bringing the cumulative installed PV capacity to 3.6 GW. In 2009, a new investment incentive of ¥ 70,000/kW for systems smaller than 10 kW, and a new surplus power purchase scheme, with a purchase price of ¥ 48/kWh for systems smaller than 10 kW, was introduced and the start of the discussion about a wider feed-in tariff.

In April 2011, METI (Ministry for Economy, Trade and Industry) announced a change in the feed-in tariffs and increased the tariff for commercial installations from ¥ 20 to 40/kWh and decreased the tariff for residential installations to ¥ 42/kWh.

As a consequence of the accident at the Fukushima Daiichi Nuclear Power Plant, Prime Minister Naoto Kan announced an overall review of the country's Basic Energy Plan. At the G8 Summit held in Deauville, France, on May 26, 2011, he announced that Japan plans to increase the share of renewable energy in total electricity supply to over 20% in the 2020s. One measure to achieve this goal is to install a PV system on some 10 million houses suitable for it.

**People's Republic of China** The 2010 Chinese PV market estimates are between 530 and 690 MW, bringing the cumulative installed capacity to about 1 GW. This is a significant increase from the 160 MW in 2009, but still only 5–7% of the total photovoltaic production. This situation will change because of the revision of the PV targets for 2015 and 2020. According to press reports, the National Energy Administration doubled its capacity target for installed photovoltaic electricity systems to 10 GW in 2015 and further up to 50 GW in 2020 [95].

According to the 12th Five-Year Plan, which was adopted on March 14, 2011, China intends to cut its carbon footprint and be more energy efficient. The targets are 17% less carbon dioxide emissions and 16% less energy consumption unit of GDP. The total investment in the power sector under the 12th Five-Year Plan is expected to reach \$803 billion (€ 618 billion), divided into \$416 billion (€ 320 billion), or 52%, for power generation, and \$386 billion

(€ 298 billion) to construct new transmission lines and other improvements to China's electrical grid.

Renewable, clean, and nuclear energy are expected to contribute to 52% of the increase and it is planned to increase power generation capacity from nonfossil fuels to 474 GW by 2015.

The investment figures necessary are in-line with a World Bank report stating that China needs an additional investment of \$64 billion (€ 49.2 billion) annually over the next 2 decades to implement an “energy-smart” growth strategy [96]. However, the reductions in fuel costs through energy savings could largely pay for the additional investment costs according to the report. At a discount rate of 10%, the annual net present value (NPV) of the fuel cost savings from 2010 to 2030 would amount to \$145 billion (€ 111.5 billion), which is about \$70 billion (€ 53.8 billion) more than the annual NPV of the additional investment costs required.

**South Korea** In 2010, about 180 MW of new PV systems were installed in South Korea, about the same as the year before, bringing the cumulative capacity to a total of 705 MW. The Korean PV industry expects a moderate increase for 2011, due to the fact that the feed-in tariff scheme is in its final year and the Korean Government continues its “One Million Green Homes” Project, as well as other energy projects in the provinces. The implementation of the Renewable Portfolio Standard in 2012 has additional consequences, as systems will need to be installed by the end of 2011 to generate electricity.

In January 2009, the Korean Government had announced the Third National Renewable Energy Plan, under which renewable energy sources will steadily increase their share of the energy mix between now and 2030. The Plan covers such areas as investment, infrastructure, technology development, and programs to promote renewable energy. The new Plan calls for a renewable energies share of 4.3% in 2015, 6.1% in 2020, and 11% in 2030.

**Taiwan** In June 2009, the Taiwan Legislative Yuan gave its final approval to the Renewable Energy Development Act, a move that is expected to bolster the development of Taiwan's green energy industry. The new law authorizes the Government to enhance

incentives for the development of renewable energy via a variety of methods, including the acquisition mechanism, incentives for demonstration projects, and the loosening of regulatory restrictions. The goal is to increase Taiwan's renewable energy generation capacity by 6.5 GW to a total of 10 GW within 20 years. In January 2011, the Ministry of Economic Affairs (MOEA) announced the revised feed-in tariffs for 2011. In 2011, the price paid by the state-owned monopoly utility, Taiwan Power, will fall 30% from 11.12 NT\$/kWh (0.264 €/kWh) to 7.33 NT\$/kWh per kWh (0.175 €/kWh) for solar installations, with an exception for rooftop installations which will be eligible for rates of 10.32 NT\$/kWh (0.246 €/kWh).

Despite the favorable feed-in tariff, the total installed capacity at the end of 2010 was only between 19 and 20 MW and the annual installation of about 7–8 MW was far less than 1% of the 3.2 GW solar cell production in Taiwan that year.

**Thailand** Thailand enacted a 15-year Renewable Energy Development Plan (REDP) in early 2009, setting the target to increase the Renewable Energy share to 20% of final energy consumption of the country in 2022. Besides a range of tax incentives, solar photovoltaic electricity systems are eligible for a feed-in premium or “Adder” for a period of 10 years. However, there is a cap of 500 MW eligible for the original 8 THB/kWh (0.182 €/kWh) “Adder” (facilities in the three Southern provinces and those replacing diesel systems are eligible for an additional 1.5 THB/kWh (0.034 €/kWh)), which was reduced to 6.5 THB/kWh (0.148 €/kWh) for those projects not approved before June 28, 2010.

As of October 2010, applications for 1.6 GW, under the *Very Small Power Producer Programme* (VSPP), and 477 MW, under the *Small Power Producer Programme* (SPP), were submitted. In 2010, it is estimated that between 20 and 30 MW were actually added, increasing the total cumulative installed capacity to 60–70 MW.

### Emerging Markets

- **Bangladesh:** In 1997, the Government of Bangladesh established the Infrastructure Development Company Limited (IDCOL) to promote economic development in Bangladesh. In

2003, IDCOL started its Solar Energy Programme to promote the dissemination of solar home systems (SHS) in the remote rural areas of Bangladesh, with the financial support from the World Bank, the Global Environment Facility (GEF), the German Kreditanstalt für Wiederaufbau (KfW), the German Technical Cooperation (GTZ), the Asian Development Bank, and the Islamic Development Bank. Since the start of the program, more than 950,000 SHS, with an estimated capacity of 39 MW, have been installed in Bangladesh by May 2011.

According to a press report, the Government plans to implement a mega project of setting up 500 MW of PV electrical power generation and the Asian Development Bank (ADB) has, in principal, agreed to provide financial support to Bangladesh for implementing the project within the framework of the Asian Solar Energy Initiative [97, 98].

- **Indonesia:** The development of renewable energy is regulated in the context of the national energy policy by Presidential Regulation No.5/2006 [99]. The decree states that 11% of the national primary energy mix in 2025 should come from renewable energy sources. The target for solar PV is 870 MW by 2024. At the end of 2010, about 20 MW of solar PV systems were installed, mainly for rural electrification purposes.
- **Malaysia:** The Malaysia Building Integrated Photovoltaic (BIPV) Technology Application Project was initiated in 2000 and at the end of 2009 a cumulative capacity of about 1 MW of grid-connected PV systems has been installed.

The Malaysian Government officially launched their GREEN Technology Policy in July 2009 to encourage and promote the use of renewable energy for Malaysia's future sustainable development. By 2015, about 1 GW must come from Renewable Energy Sources according to the Ministry of Energy, Green Technology and Water (KETHHA). The Malaysian Photovoltaic Industry Association (MPIA) proposed a five-year program to increase the share of electricity generated by photovoltaic systems to 1.5% of the national demand by 2015. This would translate into 200 MW grid-connected and 22 MW of grid systems. In the long-term beyond 2030, MPIA is calling for a 20% PV

share. Pusat Tenaga Malaysia (PTM), and its IEA international consultant, estimated that 6,500 MW power can be generated by using 40% of the nation's house rooftops (2.5 million houses) and 5% of commercial buildings alone. To realize such targets, a feed-in tariff is still under discussion and it is hoped to be under way in the second half of 2011.

First Solar (USA), Q Cells (Germany), and Sunpower (USA) have started to set up manufacturing plants in Malaysia, with a total investment of RM 12 billion and more than 2 GW of production capacities. Once fully operational, these plants will provide 11,000 jobs and Malaysia will be the world's sixth largest producer of solar cells and modules.

- *The Philippines:* The Renewable Energy Law was passed in December 2008 [100]. Under the Law, the Philippines has to double the energy derived from Renewable Energy Sources within 10 years. On June 14, 2011, Energy Secretary, Rene Almendras unveiled the new Renewable Energy Roadmap, which aims to increase the share of renewables to 50% by 2030. The program will endeavor to boost renewable energy capacity from the current 5.4 GW to 15.4 GW by 2030.

Early 2011, the country's Energy Regulator National Renewable Energy Board (NREB) has recommended a target of 100 MW of solar installations that will be constructed in the country over the next 3 years. A feed-in tariff of 17 PHP/kWh (0.283 €/kWh) was suggested, to be paid from January 2012 on. The initial period of the program is scheduled to end on December 31, 2014.

At the end of 2010, about 10 MW of PV systems were installed, mainly off-grid. SunPower has two cell manufacturing plants outside of Manila. Fab. No 1 has a nameplate capacity of 108 MW and Fab. No 2 adds another nameplate capacity of 466 MW.

- *Vietnam:* In December 2007, the National Energy Development Strategy of Vietnam was approved. It gives priority to the development of renewable energy and includes the following targets: increase the share of renewable energies from negligible to about 3% (58.6 GJ) of the total commercial primary

energy in 2010, to 5% in 2020, 8% (376.8 GJ) in 2025, and 11% (1.5 TJ) in 2050.

The Indochinese Energy Company (IC Energy) broke ground for the construction of a thin-film solar panel factory with an initial capacity of 30 MW and a final capacity of 120 MW in the central coastal Province of Quang Nam on May 14, 2011.

In March 2011, First Solar broke ground on its four-line photovoltaic module manufacturing plant (250 MW) in the Dong Nam Industrial Park near Ho Chi Minh City.

## North America

**Canada** In 2010, Canada more than tripled its cumulative installed PV capacity to about 420 MW, with 300 MW new installed systems. This development was driven by the introduction of a feed-in tariff in the Province of Ontario, enabled by the "*Bill 150, Green Energy and Green Economy Act, 2009.*" On the Federal level, only an accelerated capital cost allowance exists under the Income Tax Regulations. On a Province level, nine Canadian Provinces have *Net Metering Rules*, with solar photovoltaic electricity as one of the eligible technologies, *Sales Tax Exemptions* and *Renewable Energy Funds* exist in two Provinces and *Micro Grid Regulations* and *Minimum Purchase Prices* each exist in one Province.

The Ontario feed-in tariffs were set in 2009 and depend on the system size and type, as follows:

Rooftop or ground-mounted ≤ 10 kW	80.2 ¢/kWh (0.59 €/kWh)
Rooftop > 10 kW ≤ 250 kW	71.3 ¢/kWh (0.53 €/kWh)
Rooftop > 250 kW ≤ 500 kW	63.5 ¢/kWh (0.47 €/kWh)
Rooftop > 500 kW	53.9 ¢/kWh (0.40 €/kWh)
Ground-mounted* > 10 kW ≤ 10 MW	44.3 ¢/kWh (0.33 €/kWh)

\*Eligible for Aboriginal or community adder

The feed-in tariff scheme has a number of special rules, ranging from eligibility criteria, which limit

the installation of ground-mounted PV systems on high-yield agricultural land to domestic content requirements and *additional “price adders” for Aboriginal and community-based projects*. Details can be found in the Feed-in Tariff Program of the Ontario Power Authority [101].

**United States** With close to 900 MW of new installed PV capacity, the USA reached a cumulative PV capacity of 2.5 GW at the end of 2010. Utility PV installations more than tripled compared to 2009 and reached 242 MW in 2010. The top ten states – California, New Jersey, Nevada, Arizona, Colorado, Pennsylvania, New Mexico, Florida, North Carolina, and Texas accounted for 85% of the US grid-connected PV market [74].

PV projects with Power Purchase Agreements (PPAs), with a total capacity of 6.1 GW, are already under contract and to be completed by 2014 [102]. If one adds those 10.5 GW of projects which are already publicly announced, but PPAs have yet to be signed, this make the total “pipeline” more than 16.6 GW.

After years of political deadlock and negotiations concerning the support of renewable energies in the USA, things started to move in 2005. The main breakthrough was reached, when the 2005 Energy Bill was passed by the Senate on July 29, 2005 and signed by President Bush on August 8, 2005. The next milestone was the final approval of the Californian “Million Solar Roofs Plan” or Senate Bill 1 (SB1) in 2006. The “*Energy Improvement and Extension Act of 2008*” as part of H.R. 1424, the “*Emergency Economic Stabilization Act of 2008*” and the “*American Reinvestment and Recovery Act of 2009*” were the next steps to support the implementation of renewable energies and solar photovoltaic electricity generation.

There is no single market for PV in the USA, but a conglomeration of regional markets and special applications for which PV offers the most cost-effective solution. In 2005, the cumulative installed capacity of grid-connected PV systems surpassed that of off-grid systems. Since 2002, the grid-connected market has been growing much faster, thanks to a wide range of “buy-down” programs, sponsored either by States or utilities.

Many State and Federal policies and programs have been adopted to encourage the development of markets for PV and other renewable technologies. These consist

of direct legislative mandates (such as renewable content requirements) and financial incentives (such as tax credits). DOE has defined a financial incentive as one that: (1) transfers economic resources by the Government to the buyer or seller of goods or a service that has the effect of reducing the price paid or increasing the price received; (2) reduces the cost of producing the goods or service; and/or (3) creates or expands a market for producers [103]. Financial incentives typically involve appropriations or other public funding, whereas direct mandates typically do not. In both cases, these programs provide important market development support for PV. The types of incentives are described below. Amongst them, investment rebates, loans, and grants are the most commonly used – 42 States, the District of Columbia and the Virgin Islands – have such programs in place. Most common mechanisms are:

- Personal tax exemptions (Federal Government, 24 States + Puerto Rico)
- Corporate tax exemptions (Federal Government and 25 States)
- Sales tax exemptions for renewable investments (28 States + Puerto Rico)
- Property tax exemptions (34 States + Puerto Rico, 15 local)
- Buy-down programs (23 States + District of Columbia, Puerto Rico, Virgin Islands, 398 utilities, 17 local)
- Loan programs and grants (Federal Gov., 42 States + District of Columbia, Virgin Islands; 72 utilities, 22 local, 13 private)
- Industry support (Federal Government, 21 States + Puerto Rico, 1 local)

One of the most comprehensive databases about the different support schemes in the USA is maintained by the Solar Centre of the State University of North Carolina. The Database of State Incentives for Renewable Energy (DSIRE) is a comprehensive source of information on State, local, utility, and selected federal incentives that promote renewable energy. All different support schemes are described there and it is highly recommended to visit the DSIRE web-site <http://www.dsireusa.org/> and the corresponding interactive tables and maps for more details.

## Conclusions

The increase of conventional energy prices has increased the investment attention for renewable energies and in particular photovoltaics significantly. Despite the difficult economic conditions the worldwide production continued to increase with growth rates above 40%. There is a dynamic growth rate in system installation, but due to the tremendous capacity increase production surplus will continue to exist at least for the next 2–3 years. This development is connected to an increasing industry consolidation, which presents a risk and an opportunity at the same time. If the new large solar cell companies use their cost advantages to offer lower-priced products, customers will buy more solar systems and it is expected that the PV market will show an accelerated growth rate. However, this development will influence the competitiveness of small and medium companies as well. To survive the price pressure of the big and financial strong companies, they have to specialize in niche markets with high value added in their products. The other possibility is to offer technologically more advanced and cheaper solar cell concepts.

Even with the current economic difficulties, the increasing number of market implementation programs worldwide, as well as the overall rising energy prices and the pressure to stabilize the climate, will continue to keep the demand for solar systems high. In the long-term, growth rates for Photovoltaics will continue to be high, even if the economic frame conditions vary and can lead to a short-term slowdown. This view is shared by an increasing number of financial institutions, which are turning toward renewables as a sustainable and lucrative long-term investment. Increasing demand for energy is pushing the prices for fossil energy resources higher and higher. Already in 2007, a number of analysts predicted that oil prices could well hit 100 \$/bbl by the end of 2007 or early 2008 [104]. After the spike of oil prices in July 2008, with close to 150\$/bbl, prices have decreased due to the worldwide financial crisis and hit a low around 37 \$/bbl in December 2008. However, the oil price has rebounded and is back in the 80 \$/bbl range in the first half of 2010. It is obvious that the fundamental trend of increasing demand for oil will drive the oil price higher again. In an interview at the beginning of March 2009, the IEA

Executive Director Nobuo Tanaka warned that the next oil crisis with oil prices at around 200 \$/bbl due to a supply crunch, could be as close as 2013 because of lack of investments in new oil production.

Already for a few years, we have now observed a continuous rise of oil and energy prices, which highlights the vulnerability of our current dependence on fossil energy sources and increases the burden developing countries are facing in their struggle for future development. On the other hand, we see a continuous decrease in production costs for renewable energy technologies as a result of steep learning curves. Due to the fact that external energy costs, subsidies in conventional energies and price volatility risks are generally not taken into consideration, renewable energies and Photovoltaics are still perceived as more expensive in the market than conventional energy sources. Nevertheless, electricity production from Photovoltaic solar systems have already shown now that it can be cheaper than peak prices in the electricity exchange in a wide range of countries and if market growth continues at the current pace, electricity generation cost with Photovoltaic systems will have reached grid parity in most of Europe by 2020. In addition, solar photovoltaic generated electricity offers, contrary to conventional energy sources, a reduction of prices rather than an increase in the future.

## Future Directions

The progress of photovoltaics will depend on a parallel development of markets and progress in research. A number of science areas can make a big impact in the future.

*Material Science:* Fundamental material research and the systematic screening, synthetization, and characterization of potential solar cell materials can play an important role to find and identify new solar cell materials or substitute certain rare materials in the current family of solar cells. In the field of solar photovoltaics, the range of used materials is limited to few elements like silicon (wafer based and thin film), GaAs and its derivatives, CdTe, a few chalcopyrites ( $\text{CuInGa}(\text{SSe})_2$ ) and some dye and organic compounds. Already in the first half of the 1950s I-II-VI<sub>2</sub> compounds were researched and the first II-IV-V<sub>2</sub> compounds were synthesized. The invention of the  $\text{CuInSe}_2/\text{CdS}$  solar cell in

the early 1970s at Bell Labs spurred an increased research activity to use compound semiconductors as base material for solar cells. Fundamental theoretical bandstructure calculations were performed and identified a wide range of compound materials as possible candidates for solar cells in the 1970s. However, the systematic synthetization and investigation and characterization of these potential materials has been not done so far, but offer a chance for new material compositions and or efficiency increases.

*Microelectronics:* The further large-scale implementation of PV modules and systems will require *intelligent modules* in order minimize losses attributed to partial shading or power fluctuations. To realize this smarter control strategies and alternative power electronics topologies that dynamically optimize the yearly production of these modules have to be developed. There is a need to control (e.g., the conversion factor of DC/DC converters) and monitor (e.g., distributed temperature sensing) various parameters in real time to enable the plant-level controller to optimize the energy yield. Ultimately, he might be able to make trade-offs between lifetime and maximizing power here and now. From a technological point of view, this implies that additional power electronic circuits and sensors need to be placed in and around the module.

*Storage Technologies:* Future grid-connected PV systems will be subject to more stringent regulatory requirements for the delivery of “ancillary services” to support the electricity grid when reserve and reactive power injection (for voltage support) has to be delivered. As the electricity grid has to deal with positive as well as negative balances, this involves the “shaving” of peak production and temporarily boosting power output. Lowering the output is easily achieved by moving away from the MPP, but when storage is at hand, the energy conversion can be kept at maximum level and the output difference is stored for later recovery.

Such storage functions may be centralized or distributed – a possibly micro-storage for short-term needs could be introduced at the module level in close conjunction to DC/DC converters. These storage components could consist of improved supercapacitors with low leakage and innovative thin-film battery approaches.

To realize such innovative approaches needs the further development of the respective power

components and storage technologies as solar modules increase in temperature during operation which is not favorable for the lifetime of current power electronics and storage technologies.

### Exchange Rates

1 € = 1.35 CAD  
 1 € = 25.52 CZK  
 1 € = 9.5 RMB  
 1 € = 60 INR  
 1 € = 1,450 KRW  
 1 € = 42 NT\$  
 1 € = 60 PHP  
 1 € = 44 THB  
 1 € = 1.30 US\$

### Bibliography

1. Paula Mints (2010) Manufacturer shipments, capacity and competitive analysis 2009/2010. Navigant Consulting Photovoltaic Service Program, Palo Alto
2. Paula Mints, Global PV demand 2011 and beyond, Webinar: vote solar, 12 Jan 2011
3. PV News, published by Greentech Media (2011) ISSN 0739-4829
4. Photon International, March 2011
5. World Economic Forum (2011) Green investing 2011 – reducing the cost of financing, Ref. 200311, April 2011
6. The PEW Charitable Trusts (2011) Who's winning the clean energy race? 2010 edition. [www.pewtrusts.org](http://www.pewtrusts.org)
7. Photon Photovoltaic Stock Index PPVX, [http://www.photon-international.com/news/news\\_01-08-01\\_eu\\_ppvx.htm](http://www.photon-international.com/news/news_01-08-01_eu_ppvx.htm)
8. IMS Research, Press release, 31 May 2011
9. iSupply, PV perspectives, Feb 2011
10. Homan G, Presentation at intersolar 2009
11. Bernreuther J, Haugwitz F (2010) The who's who of silicon production. Bernreuter Reserach, Würzburg, Germany
12. Green M, Emery K, Hishikawa Y, Warta W, Dunlop EE (2011) Solar cell efficiency tables (version 38). Prog Photovolt: Res Appl 19:565–572
13. Zhao J, Wang A, Green MA, Ferrazza F (1998) Novel 19.8% efficient “honeycomb” textured multicrystalline and 24.4% monocrystalline silicon solar cells. Appl Phys Lett 73:1991–1993
14. Schultz O, Glunz SW, Willeke GP (2004) Multicrystalline silicon solar cells exceeding 20% efficiency. Prog Photovolt: Res Appl 12:553–558
15. Shockley W, Queisser HJ (1961) Detailed balance limit of efficiency of p-n junction solar cells. J Appl Phys 32(3):510–519
16. Swanson RM (2005) Approaching the 29% limit efficiency of silicon solar cells. In: Proceedings of the 20th European photovoltaic solar energy conference, Dresden, 584–589 pp. ISBN 3-936338-25-6

17. Taguchi M, Tsunomura Y, Inoue H, Taira S, Nakashima T, Baba T, Sakata H, Maruyama E, (2009) High-efficiency HIT solar cell on thin (<100 μm) silicon wafer. In: Proceedings of the 24th European photovoltaic solar energy conference, Hamburg, 1690–1693 pp. ISBN 3-936338-25-6
18. Swanson R (2008) The SunPower story: The path from R&D concentrator cells to a high volume PV panel and system manufacturer. In: Proceedings of the 33 rd IEEE photovoltaic specialists conference, San Diego
19. Carlson DE, Wronski CR (1976) Amorphous silicon solar-cell. *Appl Phys Lett* 28(11):671–673
20. Staebler DL, Wronski CR (1977) Reversible conductivity changes in discharge-produced amorphous Si. *Appl Phys Lett* 31(4):292–294
21. Benagli S, Borrello D, Vallat-Sauvain E, Meier J, Kroll U, Hoetzel J, Bailat J, Steinhauser J, Marmelo M, Monteduro M, Castens L (2009) High-efficiency amorphous silicon devices on LPCVD-ZnO-TCO prepared in industrial KaiTM-M R&D reactor. In: Proceedings of the 24th European photovoltaic solar energy conference, Hamburg, 2293–2298 pp. ISBN 3-936338-25-6
22. Yang J, Guha S (1992) Double-junction amorphous silicon-based solar-cells with 11-percent stable efficiency. *Appl Phys Lett* 61(24):2917–2919
23. Yamamoto K, Nakashima A, Suzuki T, Yoshimi M, Nishio H, Izumina M (1994) Thin-film polycrystalline Si solar cell on glass substrate fabricated by a novel Low temperature process. *Jpn J Appl Phys* 33:L1751–L1754
24. Meier J, Dubail S, Platz R, Torres P, Kroll U, Selvan JA, Pellaton Vaucher N, Hof C, Fischer D, Keppner H, Flückiger R, Shah A, Shklover S, Ufert K-D (1997) Towards high-efficiency thin-film silicon solar cells with the “micromorph” concept. *Sol Energy Mat Sol C* 49(1–4):35–44
25. Wu X, Keane JC, Dhare RG, DeHart C, Albin DS, Duda A, Gessert TA, Asher S, Levi DH, Sheldon P (2001) In: Proceedings of the 17th European photovoltaic solar energy conference 2001. ISBN 3-936338-08-6, p 995
26. Goncalves LM, Bermudez VD, Ribeiro HA, Mendes AM (2008) Dye-sensitized solar cells: a safe bet for the future. *Energ Environ Sci* 1(6):655–667
27. Sinha P, Kriegner ChJ, Schew WA, Kaczmar SW, Traister M, Wilson DJ (2008) Regulatory policy governing cadmium-telluride photovoltaics: a case study contrasting life cycle management with the precautionary principle. *Energ Policy* 36:381–387
28. Zayed J, Philippe S (2009) Acute oral and inhalation toxicities in rats with cadmium telluride. *Int J Toxicol* 28(4):259–265
29. Wadia C, Alivisatos AP, Kammen DM (2009) Materials availability expands the opportunity for large-scale photovoltaic deployment. *Environ Sci Technol* 43(6):2072–2077
30. Wagner S, Shay JL, Migliorato P, Kasper HM (1974) CuInSe<sub>2</sub>/CdS heterojunction photovoltaic detectors. *Appl Phys Lett* 25:434
31. Dimmler B, Schock HW (1996) Scaling-up of CIS technology for thin-film solar modules. *Prog Photovolt: Res Appl* 4(6):425–433
32. Jackson P, Hariskos D, Lotter E, Paetel S, Wuerz R, Menner R, Wischmann W, Powalla M (2011) New world Record efficiency for Cu(In,Ga)Se<sub>2</sub> thin-film solar cells beyond 20%. *Prog Photovolt: Res Appl*. doi: 10.1002/pip.1078
33. Kushiya K (2009) Key near-term R&D issues for continuous improvement in CIS-based thin-film PV modules. *Sol Energy Mat Sol C* 93:1037–1041
34. Meeder A, Neisser A, Rühle U, Mayer N (2007) Manufacturing the first MW of large-area CuInS<sub>2</sub>-based solar modules – recent experiences and progress. In: Proceedings of the 22nd European photovoltaic solar energy conference, Milan, p 2115. ISBN 3-936338-22-1
35. Bosi M, Pelosi C (2007) The potential of III-V semiconductors as terrestrial photovoltaic devices. *Prog Photovolt: Res Appl* 15(1):51–68
36. King RR, Boca A, Hong W, Liu X-Q, Bhusari D, Larrabee D, Edmondson KM, Law DC, Fetzer CM, Mesropian S, Karam NH (2009) Band-gap-engineered architectures for high-efficiency multijunction concentrator solar cells. In: Proceedings of the 24th European photovoltaic solar energy conference, Hamburg. ISBN 3-936338-25-6
37. Gratzel M (2001) Photoelectrochemical cells. *Nature* 414(6861):338–344
38. O'Regan B, Gratzel M (1991) A Low-cost, high-efficiency solar-cell based on dye-sensitized colloidal TiO<sub>2</sub> films. *Nature* 353(6346):737–740
39. Li G, Shrotriya V, Huang JS, Yao Y, Moriarty T, Emery K, Yang Y (2005) High-efficiency solution processable polymer photovoltaic cells by self-organization of polymer blends. *Nat Mater* 4(11):864–868
40. Brabec CJ (2004) Organic photovoltaics: technology and market. *Sol Energy Mat Sol C* 83(2–3):273–292
41. Krebs FC (2005) Alternative PV: large scale organic photovoltaics. *REfocus* 6(3):38–39
42. U.S. Photovoltaic Industry Roadmap Steering Committee (2001) Solar-electric power: the U.S. photovoltaic industry roadmap, 36 p. Solar Electricity Industry Association, Washington, DC, USA
43. Navigant Consulting Inc (2006) A review of PV inverter technology cost and performance projections. National Renewable Energy Laboratory, Golden, 100 pp
44. EU PV European Photovoltaic Technology Platform (2007) A strategic research agenda for photovoltaic solar energy technology. European Communities. Sixth European Framework Programme for research and technological development, Luxembourg, 76 p. ISBN 978-92-79-05523-2
45. Kroposki B, Margolis R, Kuswa G, Torres J, Bower W, Key T, Ton D (2008) Renewable systems interconnection. National Renewable Energy Laboratory, Golden, Colorado, 23 p
46. NEDO (2009) The roadmap PV2030+; new energy and industrial technology organization (NEDO), Kawasaki, Japan
47. Kurokawa K, Aratani F (2004) Perceived technical issues accompanying large PV development and Japanese



- "PV2030". In: Proceedings of the 19th European photovoltaic solar energy conference and exhibition, Paris. ISBN 3-936338-14-0
48. Government of China (2006) Medium and long-term planning for scientific and technological development (2006–2020), Feb 2006
  49. National Development and Reform Commission (2007) Medium and long-term development plan for renewable energy in China, Sep 2007
  50. Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions A European Strategic Energy Technology Plan (SET-Plan) – Towards a low carbon future; 22 November 2007; COM(2007) 723 final, {SEC(2007) 1508}; {SEC(2007) 1509}; {SEC(2007) 1510}; {SEC(2007) 1511}
  51. Ministry of New & Renewable Energy, Government of India Solar mission document (2009) <http://mnre.gov.in/pdf/mission-document-JNSM.pdf>
  52. Indian Ministry of New & Renewable Energy Thrust Areas of R & D in solar photovoltaic technology (2008) <http://www.mnre.gov.in/spv-thrust.htm>
  53. NEDO Brochure, Energy and Environment Technologies, December 2007. [http://www.nedo.go.jp/kankobutsu/pamphlets/kouhou/2007gaiyo\\_e/87\\_140.pdf](http://www.nedo.go.jp/kankobutsu/pamphlets/kouhou/2007gaiyo_e/87_140.pdf)
  54. U.S. Department of Energy (2008) Solar energy technologies programme (solar programme): 2008–2012 multi-year programme plan, Apr 2008
  55. Company Web-sites and Press Releases
  56. Data received from Companies during personal visits
  57. Bloomberg (2010) Bloomberg new energy finance – renewable energy data. <http://bnef.com/>
  58. Yang CJ (2010) Reconsidering solar grid parity. *Energy Policy* 38:3270–3273
  59. Maycock PD (2002) The world photovoltaic market – report (January). PV Energy Systems, Warrenton
  60. Parente V, Goldemberg J, Zilles R (2002) Comments on experience curves for PV modules. *Prog Photovolt: Res Appl* 10:571–574
  61. Neij L (2008) Cost development of future technologies for power generation—a study based on experience curves and complementary bottom-up assessments. *Energy Policy* 36:2200–2211
  62. International Energy Agency (2010) Technology roadmap. Solar Photovoltaic Energy. International Energy Agency, Paris, France
  63. Hoffmann W (2009) The role of PV solar electricity to power the 21st century's global prime energy demand. *IOP Conf Ser: Earth Environ Sci* 8:012007
  64. Hoffmann W, Wieder S, Pellkofer T (2009) Differentiated price experience curves as evaluation tool for judging the further development of crystalline silicon and thin film PV solar electricity products. In: 24th European photovoltaic solar energy conference, Hamburg, Germany, 21–25 Sep 2009
  65. Jäger-Waldau A (2010) Status and perspectives of thin film photovoltaics. In: *Thin film solar cells: current status and future trends*. Nova Publishers, New York
  66. Maycock P (1976–2003) PV News. PV Energy Systems
  67. Schaeffer GJ, Seebregts AJ, Beurskens LWM, Moor HHC, Alsema EA, Sark W, Durstewicz M, Perrin M, Boulanger P, Laukamp H, Zuccaro C (2004) Learning from the sun: analysis of the use of experience curves for energy policy purposes – the case of photovoltaic power. Final Report of the Photex Project, DEGO: ECN-C-04-035
  68. Hoff TE, Pasquier BJ, Peterson JM (2010) Market transformation benefits of a PV incentive program. In: SOLAR 2010 conference proceedings, American Solar Energy Society, Phoenix, Arizona, 17–22 May 2010
  69. Barbose G, Darghouth N, Wiser R (2010) Tracking the Sun III: the installed cost of photovoltaics in the U.S. from 1998–2009. Lawrence Berkeley National Laboratory, Berkeley
  70. Urbschat C, Barban F, Baumgartner B, Beste M, Herr M, Schmid-Kieninger A, Rossani F, Stry-Hipp G, Welke M (2002) Sunrise 2002 – the solar thermal and photovoltaic markets in Europe, Berlin
  71. Jäger-Waldau A (2005) Photovoltaics status report 2005: research, solar cell production and market implementation of photovoltaics. Office for Official Publications of the European Communities, EUR 21836 EN. ISBN 92-79-00174-4
  72. Bundesverband Solarwirtschaft e.V (2010) Statistische Zahlen der deutschen Solarstrombranche (photovoltaik). Bundesverband Solarwirtschaft e.V. (BSW Solar), Berlin, Germany, 4 pp
  73. Solar Energy Industry Association (SEIA) (2010) U.S. solar market insight, 2nd Quarter 2010, executive summary
  74. Solar Energy Industry Association (SEIA) (2011) U.S. solar market insight, 1st Quarter 2011, executive summary
  75. Bundesverband Solarwirtschaft e.V (2011) Statistische Zahlen der deutschen Solarstrombranche (photovoltaik). Bundesverband Solarwirtschaft e.V. (BSW Solar), Berlin, Germany, 4 pp
  76. Jäger-Waldau A PV status report 2011, Office for Official Publications of the European Union, EUR 24807 EN. ISBN 978-92-79-20171-4
  77. ICIS news, Asia polysilicon prices to firm in 2011 on solar demand, 13 Jan 2011
  78. Semi PV Group, Semi China Advisory Committee and China PV Industry Alliance (CPIA) (2011) China's solar future – a recommended China PV policy roadmap 2.0, April 2011. Semiconductor Industry Association, San Jose, USA
  79. Ikki O (2011) PV activities in Japan. PV status report, vol 17, no. 5, May 2011
  80. European Photovoltaic Industry Association (2011) Global market outlook for photovoltaics until 2015. European Photovoltaic Industry Association, Brussels, Belgium
  81. Photovoltaic Energy Barometer (2011) Systèmes Solaires, le journal du photovoltaïque n° 5 – 2011, April 2011, ISSN 0295–5873
  82. German Federal Network Agency (Bundesnetzagentur), Press release 21 Mar 2011

83. Gestore Servizi Energetici, Press release, 15 Feb 2011
84. European Wind Energy Association (2011) Wind in power – 2010 European statistics, Feb 2011
85. Platts (2011) Power in Europe, Jan 2011. [www.platts.com](http://www.platts.com)
86. Ministère de l'Économie, de l'industrie et de l'emploi, Press release, 24 Feb 2011
87. Gesetz über den Vorrang Erneuerbaren Energien (Erneuerbare-Energien-Gesetz – EEG), Bundesgesetzblatt Jahrgang 2000 Teil I, Nr. 13, p 305 (29.03.2000)
88. German Federal Network Agency (Bundesnetzagentur), Press release 16 June 2011
89. Gestore Servizi Energetici, Aggiornamento, 16 Aug 2011
90. Royal Decree 1565/10, published on 23 Nov 2010. <http://www.boe.es/boe/dias/2010/11/23/pdfs/BOE-A-2010-17976.pdf>
91. Royal Decree RD-L 14/10, published on 24 Dec 2010. <http://www.boe.es/boe/dias/2010/12/24/pdfs/BOE-A-2010-19757.pdf>
92. Asociación de la Industria Fotovoltaica (ASIF). <http://www.asif.org/principal.php?idseccion=565>
93. Asian Development Bank (2011) Asia solar energy initiative: a primer. ISBN 978-92-9092-314-5, April 2011. Asian Development Bank, Manila, Philippines
94. Watt M, Passey R, Johnston W (2011) PV in Australia 2010 – Australian PV survey report 2010, Australian PV Association, May 2011
95. Reuters (2011) China doubles solar power target to 10 GW by 2015, 6 May 2011. <http://www.reuters.com/article/2011/05/06/china-solar-idUKL3E7G554620110506>
96. The World Bank (2010) Winds of change – East Asia's sustainable energy future, May 2010
97. The Daily Star, Target 500 MW solar project, 15 May 2011. <http://www.thedailystar.net/newDesign/news-details.php?nid=185717>
98. UNB connect, ADB assures fund for 500 MW solar system, 4 June 2011. <http://www.unbconnect.com/component/news/task-show/id-49440>
99. Presidential Regulation 5/2006, National Energy Policy, published 25 Jan 2006
100. Republic of the Philippines, Congress of the Philippines, Republic Act No. 9513 December 16, 2008, AN ACT PROMOTING THE DEVELOPMENT, UTILIZATION AND COMMERCIALIZATION OF RENEWABLE ENERGY RESOURCES AND FOR OTHER PURPOSES
101. Ontario Power Authority, Feed-In Tariff Programme, 30 September 2009. [http://fit.powerauthority.on.ca/Storage/97/10759\\_FIT-Program-Overview\\_v1.1.pdf](http://fit.powerauthority.on.ca/Storage/97/10759_FIT-Program-Overview_v1.1.pdf)
102. Greentech Media Inc., Enfinity America Corporation (2011) The US PV Market in 2011 – Whitepaper
103. Gielecki M, Mayes F, Prete L EIA, Renewable energy 2000: issues and trends, section on Incentives, mandates, and government programs for promoting renewable energy, DOE/Energy Information Administration, DOE/EIA-0628
104. International Herald Tribune, 24 July 2007. <http://www.ihf.com/articles/2007/07/24/bloomberg/bxoil.php>

## Suggested Readings and Internet Sites

- European Commission, DG JRC, Annual PV Status Report (since 2002) <http://re.jrc.ec.europa.eu/refsys/>
- European Photovoltaic Industry Association, (various publications) [www.epia.org](http://www.epia.org)
- European Renewable Energy Council, [www.erec.org](http://www.erec.org)
- Intergovernmental Panel on Climate Change (IPCC), [www.ipcc.ch](http://www.ipcc.ch). Special report on renewable energy sources and climate change mitigation (SRREN), <http://srren.ipcc-wg3.de/report>
- International Energy Agency, photovoltaic implementation agreement <http://www.iea-pvps.org/>
- Renewable energy policy network for the 21st century, [www.ren21.net](http://www.ren21.net)
- Solar Energy Industry Association (SEIA), <http://www.seia.org/>
- United Nations Environment Programme & Bloomberg New Energy Finance, Global Trends in renewable Energy Investment (series), [www.unep.org/](http://www.unep.org/); [www.bnef.com](http://www.bnef.com)

---

## Pig Breeding for Increased Sustainability

PIETER W. KNAP

PIC International Group, Schleswig, Germany

### Article Outline

- Glossary
- Definition of the Subject
- Introduction
- Biodiversity
- Pollution
- Animal Welfare
- Future Directions
- Acknowledgments
- Bibliography

### Glossary

**Allele** Allele is one of the several possible forms of the DNA sequence at a particular locus.

**BLUP (best linear unbiased prediction)** BLUP is a method to estimate breeding values by taking account of the pedigree relationships among the individuals instead of assuming uncorrelated residuals as in ordinary least squares methodology.

**Corticosteroids** Corticosteroids are steroid hormones such as cortisol, produced in the adrenal cortex. They are involved in energy metabolism, immune response, stress response, and many other functions.

**Cryoconservation** Cryoconservation is the conservation of living material by freezing.

**Effective population size** Effective population size is a measure of the genetic variability of a population. It is defined as the number of breeding individuals in a stable population with a 1:1 sex ratio, no overlapping generations, and random mating and reproduction, that would lead to the same rate of inbreeding as what occurs in the population under study.

**Fitness constraint** Fitness constraint is a reduction of the health and strength of an individual.

**Microsatellite** Microsatellite is a repeating sequence of a few DNA base pairs, highly variable and therefore used as molecular markers in genetic analyses. Typically neutral, i.e., not associated with functional genes. The alleles differ in terms of the number of repeats, so there can be many.

**SNP (single nucleotide polymorphism)** SNP is a mutation of a single DNA base pair at a specific locus, possibly in a functional gene. SNPs are used as molecular markers in genetic analyses. They usually carry two alleles.

**Stereotypy** Stereotypy is a repetitive, apparently purposeless, behavior such as pacing, rocking, chewing, and licking during psychological distress. This behavior is seen in captive animals, often caused by the lack of options to exercise *other* instinctive behavior patterns because their required substrate (e.g., space or companions) is not available.

## Definition of the Subject

The sustainability of farm animal production depends largely on strategies for animal management, health care, and nutrition, and on strategies for processing and marketing the products (e.g., meat, eggs, milk, and manure). Strategies for animal *breeding* exploit genetic and reproductive technology to better match the next generation of production animals to what the market requires. Breeding creates gradual changes in the animal species, providing a way to support

sustainable development: better match the next generation of production animals to what enhanced sustainability requires. The technological challenge is to consider the balance among the various sustainability elements (profitability, human nutrition, environmental load, resource management, animal welfare), and to design genetic strategies to support that balance.

## Introduction

Gamborg and Sandøe [1, 2] write about “applying the notion of sustainability” in animal breeding. They first explain why this is relevant at all: “animal breeding [...] is a largely unnoticed, yet economically vital part of the agriculture and food sector. But despite remarkable advances in productivity [it has] negative impacts: for example, on animal health and welfare, and on genetic diversity. This raises the question of what limits to acceptable practice we should set in this area.” This is about societal regulation of industry practice: about a *license to breed*.

These authors notice that “discussions of sustainability may open up dialogue on ethical issues and [may] help to set an agenda,” and describe projects where “breeders were required [...] to develop a definition of sustainable farm animal breeding [...]. They were asked to identify their key concerns and priorities, to characterize any resulting dilemmas, and to suggest ways towards a meaningful operationalization.

The [...] breeders [could describe and clarify] the concerns they considered relevant [...]. But they found it harder to identify the concerns they had chosen to *exclude* [...]. Most difficult [...] was the prioritization of potentially conflicting concerns. Roughly speaking, there are two ways to overcome conflict here:

1. technological solutions, in which relevant conflict is resolved through technological changes in breeding practice
2. increased transparency, in which clear statements about the relative priorities [...] are essential.”

The rest of the above text focuses on point (2).

By contrast, the present chapter deals mainly with point (1), focusing on pig breeding while borrowing from poultry and cattle breeding where relevant.

Prioritization of the relevant, and potentially conflicting, concerns is indeed difficult, but essential in any concrete case. Such a prioritization cannot be attempted here, as it will always depend on that concrete case.

The classical “triple bottom line” of Elkington [3] is extended here with a fourth element, leading to the sustainability targets

*People – Pigs – Planet – Profit*

The possible contribution of the technology of pig breeding and genetics to two of these targets is discussed here: *Pigs* is about animal welfare, *Planet* deals with biodiversity and pollution.

*People* is about social justice, with little connection to pig breeding technology. A possible case would be biopiracy which is more a political and economic issue than a technological one, and more relevant in the plant breeding sector – but see [4] for a case study of the immigration of Meishan into western pig populations, a commercial failure due to unforeseen technological developments. Influencing *Profit* by pig breeding has been covered intensely since selection indexes were designed – there is no need to repeat that here.

It must be borne in mind throughout that “sustainability will always be a matter of more or less: it can never be an absolute goal” [2].

## Biodiversity

FAO’s *State of the world’s animal genetic resources for food and agriculture* [5] mentions 140 known extinct and 599 non-extinct pig breeds. Of the 599, 90% are *local breeds* (LB, occurring in one country only) and 6% are *international transboundary breeds* (ITB). Of these 599 breeds, 22% are “at risk” based on population size: roughly, populations with less than 1,000 breeding females or 20 breeding males are considered to be at risk, genetic survival being endangered. The risk status of 38% of the breeds is unknown, 40% are “not at risk.”

Livestock breeds become endangered in many ways. FAO [5] gives three categories: (1) emergencies: drought, flooding, earthquakes, famine, war; (2) epidemics and zoonosis eradication campaigns; (3) most important: livestock sector trends, described earlier [6–8] in terms of displacement by other breeds,

indiscriminate crossbreeding with exotic germplasm, overfocus on a single trait, no sustained breeding program, changes in production systems or producer preferences, and technology development. To this can be added the reduction of demand for the breed’s products – for pigs, since about 1950, such a product would typically be fat. Most pig breeds have evolved as an intrinsic part of a production system, catering for demands for particular products. When the production system (or the demand for its products) disappears, its associated breeds (e.g., lard-type pigs) will disappear with it – unless they find an alternative niche.

Extinction of a livestock breed can be undesirable for several reasons. A pressing one is (A) when it contributes to “maintaining the identity of human communities” [9], e.g., pigs in the South Pacific, or even more when the livelihood of a human group depends on it; this typically involves ruminant breeds. Simianer [10] gives two other categories: (B) “the insurance argument”: “genetic diversity can be seen as an insurance against future changes, [with] the objective [...] to maintain sufficient genetic diversity to be able to adapt to the challenges that are ahead” – bearing in mind that those challenges are increasingly unpredictable in times of global climatic change and intensifying global trade; and (C) cultural arguments: “farm animal breeds must be seen as a man-made good with a long history, often parallel with [human] cultural development [...] and therefore similar arguments for conservation apply as for other cultural assets [such as historical] buildings or artwork”.

All these arguments call for genetic conservation, most pressingly with regard to category (A). With regard to category (C), the question is how to objectivize the cultural *value* of a livestock breed [11], which will be necessary when the costs of its conservation are being budgeted.

With regard to category (B), the main question is what kind of future challenges might require the breeding sector to exploit genetic material that previously has been unsatisfactory enough to become endangered; why [12] “preserve animals that farmers have abandoned”? The interest must then be in traits that were not important in mainstream breeding, e.g., meat quality traits (covering unexpected changes in consumer preferences) or robustness traits that facilitate

adaptation to previously uncommon conditions, covering unexpected changes in options for health or climate control, or nutrition.

There is serious doubt among geneticists whether this insurance argument for breed conservation is realistic at all. To quote [13, 14], rearranged for consistency: “little use is made of conserved populations in mainstream commercial production of livestock; modern populations [...] are so far ahead of conserved strains in production traits, that adaptation of [the modern populations] offers far more opportunity than crossing back to far out of date stocks; even in countries where highly adapted breeds have evolved, typically these breeds are not perceived as having sufficient immediate utility to make them commercially viable; there are no present-day commercial animal-improvement companies or organizations that feel the need to invest in conservation as an insurance; the main perceived insurance benefit [...] is the conservation of adapted [sets of] alleles that are confined to one or a few breeds; there is [...] no reason to assume that there is [...] little variation in fitness associated traits in livestock populations simply due to selection; another important justification for conservation [...] is the value [...] for the increasingly powerful genomics analyses that have the potential to shed much light on the basic biology of adaptive traits; with moves towards genomic selection [...] the emphasis moves more towards best utilization of the large amounts of variation present in [...] commercial populations.” Much earlier, Dempfle [15] wrote “only in very exceptional cases would a geneticist interested in improving [a leading] breed consider going to another breed to exploit interbreed variation, since most likely this would result in lowering of the mean. This is not likely to change very much, even with future technology.”

Despite such skeptical points of view, endangered breed conservation is a current political commitment and an actively promoted reality, so it is valid to consider its technical issues.

Category (B) above centers around disappearance of (possibly) useful alleles – emphasizing within-species genetic diversity rather than particular breeds. From a technological point of view, *in vitro* cryoconservation would be adequate. So, allowing for the 38% “unknown” risk status breeds mentioned above, the technological challenge is to conserve the

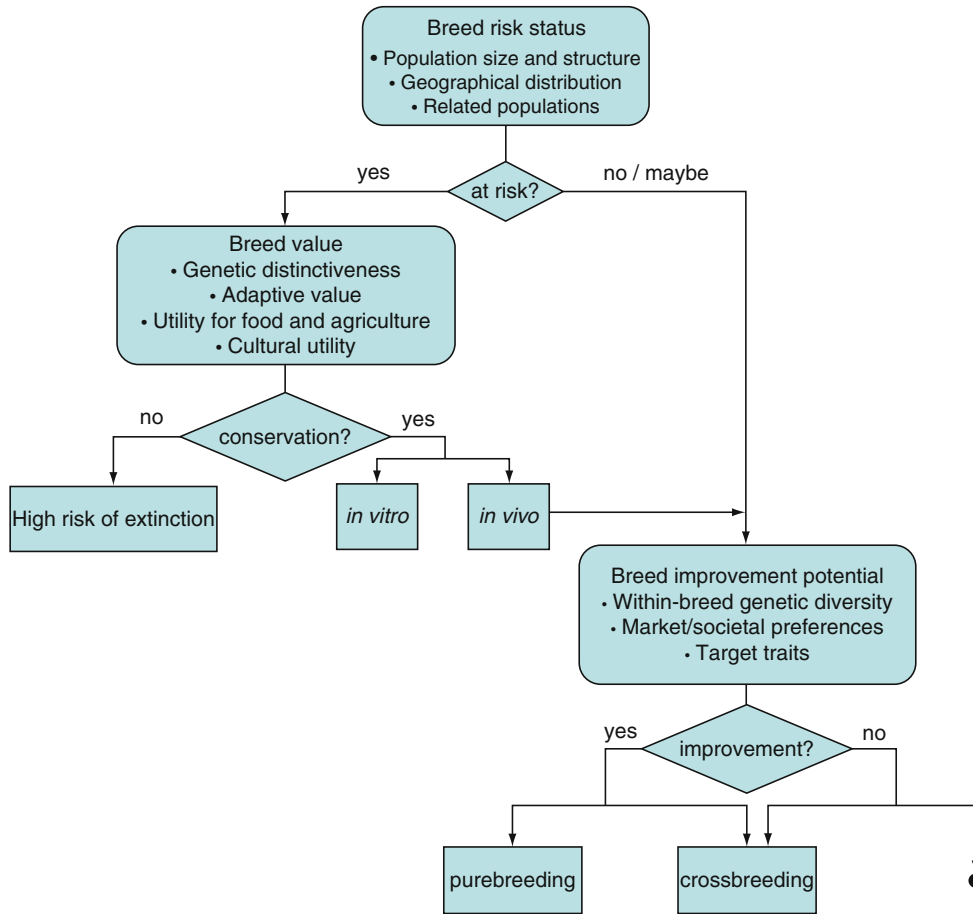
useful alleles carried by 130–360 pig breeds. Most of these are LBs.

Conservation of animal genetic resources is expensive, logistically complicated, and its worldwide funding is limited and fragmentary. Conservation of all genetic diversity is therefore not feasible and choices will have to be made with regard to funding the conservation of particular breeds, and not other ones: *global genetic resource management*. FAO [5] presents criteria (the breed’s status, value, and potential for improvement) to support decision-making around conservation and genetic improvement actions, see Fig. 1. It holds several items where animal breeding and genetics technology can usefully contribute: (1) genetic distinctiveness; (2) population size and structure; (3) utility for food and agriculture, including adaptive traits. On a higher level, the balance between distinctiveness, risk status, and utility: (4) the ultimate priority level of the breed. Further elements are (5) target traits for genetic improvement and (6) genetic improvement programs. Sections “Genetic Distinctiveness” to “Genetic Improvement Programs” deal with these items in this order.

### Genetic Distinctiveness

Megens et al. [16] describe the genetic diversity among 46 Chinese and 51 European pig breeds. Microsatellite marker genotypes were converted to genetic distance estimates among breeds, which can be worked into phylogenetic trees (dendrograms) and other cluster representations; see [17] for background information about various techniques. The results of this analysis are in Fig. 2, which shows genetic distance estimates among these breeds through three-dimensional scaling. This reveals a marked difference between the Chinese breeds (strongly diverse in all three dimensions, falling apart into five geographic clusters) versus the European breeds which form a separate cluster, much tighter when compared on the same scale. The European dendrogram (not shown here) gives more detail, suggesting distinct groups of English and south-European LBs. The authors relate their results in beautiful detail to the domestication history of the breeds.

An earlier study of roughly these same European DNA samples [18, 19] measured the *allelic richness* of these breeds, in terms of the mean effective number of



**Pig Breeding for Increased Sustainability. Figure 1**

Information required to design strategies for global genetic resource management (Modified from [5])

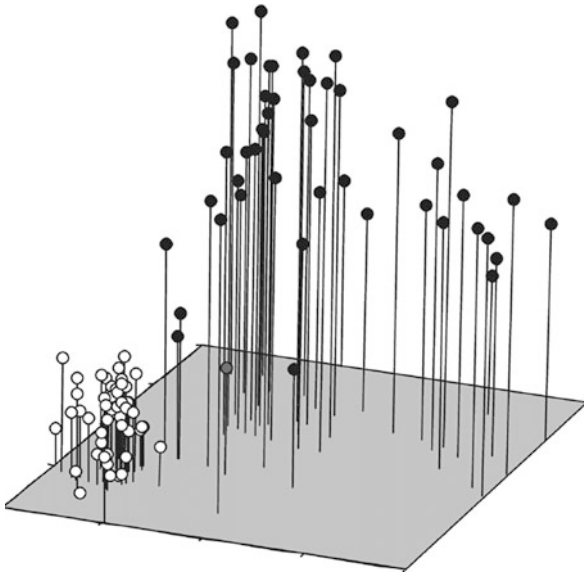
alleles per marker (2.3–3.5 in the ITBs, 1.9–4.0 in the LBs, for microsatellites), and the number of *private alleles* (found in one breed only) – zero to five, with an outlier at 15.

Many similar analyses have been reported since 1995 ([20]; four Belgian breeds). Recent ones involve three Brazilian LBs, one ITB and an industrial synthetic [21], and twelve North and South American LBs, three ITBs, and three wild varieties [22].

Obviously, the accuracy of such analyses depends on the number of markers per breed. Alves et al. [23] show a diminishing-returns pattern for the accuracy of clustering six pig breeds and wild boar, as a function of numbers of microsatellites analyzed – with ~24 markers required for 95% overall clustering accuracy, and ~35 to cluster the “less divergent populations.” The accuracy of clustering parameter

estimates is commonly quantified by bootstrapping. Felsenstein [24] notices that such resampling techniques are particularly valuable when the function to be estimated (e.g., a dendrogram) is algebraically complicated so that its variance cannot be derived analytically. An example of analytical derivation is in [25] where the standard error of the genetic distance estimate was obtained from observed allele frequencies. Contrary to data-specific procedures such as the bootstrap, analytical forms allow for algebraic rearrangement of terms so that data volume and structure required for a particular accuracy level can be obtained – as done empirically in [23].

To support the decision-making process of Fig. 1, quantification of genetic distinctiveness of a breed must consider within-breed and between-breeds diversity. The breed’s (possibly unique) alleles will be of



**Pig Breeding for Increased Sustainability. Figure 2** Estimated genetic distances among 46 Chinese (*black circles*) and 51 European (*white circles*) pig breeds, and one Sino-European synthetic (*gray circle*). The genetic distance between two breeds is quantified here by the physical distance between their two circles (From [16])

future interest more likely (1) when it shows a clear genetic distance to other breeds in the between-breeds diversity as in Fig. 2, and (2) when it shows a larger within-breed diversity (carries a larger number of different alleles). Also, a successful breeding program will be easier to implement for breeds with larger within-breed diversity, see the section on “[Genetic Improvement Programs](#)”. Therefore a breed’s genetic distinctiveness should be quantified by some combination of (1) genetic distance to other breeds and (2) within-breed genetic variability. Their relative weighting will depend on the value of parameters such as the proportion of total diversity due to diversity between breeds (Wright’s  $F_{ST}$ , estimated at  $<0.3$  in the above studies, naturally dependent on the sample’s breed composition); the “standard” approach [26] is to weight item (1) by  $F_{ST}$  and (2) by  $(1 - F_{ST})$ . It will also depend on strategic issues like the breed’s intended purpose. For example, the between-breeds component would count more if the breed must play a role in a terminal cross-breeding program [27] and less if the breed must be

merged into a synthetic for subsequent genetic improvement. Approaches are described, discussed, and/or illustrated by [28–31] and sources referenced therein. A practical suggestion [32] is “to consider how much diversity the breed adds to a core set constituted by commercial lines or breeds that are already subject to successful conservation.”

All the above presumes a random approach: no specific traits are targeted and genetic diversity is valued as a neutral entity. Neutral markers like microsatellites are appropriate if future needs are indeed unknown: “since we need to maintain the genetic capacity to cope with challenges not even known today, this can be best accomplished by maintaining neutral genetic diversity” [10].

Nevertheless, some studies have deliberately used markers associated with specific traits. Ciobanu et al. [33] argue that “the relationship between variability at neutral marker loci [...] and adaptation or individual fitness is still unclear” and “to characterize a breed not only in terms of genetic distance [...] but also in terms of variation at interesting loci associated with phenotypes, [...] will give more opportunity to elaborate an efficient strategy for conservation of breeds, maintaining their ‘useful’ genetic diversity and providing important resources for possible new unique traits.” Likewise, half of the markers typed by Iannuccelli et al. [34] were SNPs “chosen for their position close to interesting QTLs.” They mention that in addition to comparisons between breeds, they have “focused on the diversity within genomic regions containing genes that influence economically important traits, the variability of which is supposed to have evolved under the influence of artificial selection. [This has] allowed us to reveal regions where artificial selection favored certain alleles. For some SNPs where both alleles were found in all breeds, the frequencies were very different between breeds, suggesting different selection histories” (translated). A method to combine neutral diversity and diversity due to selection into a single criterion is presented in [35].

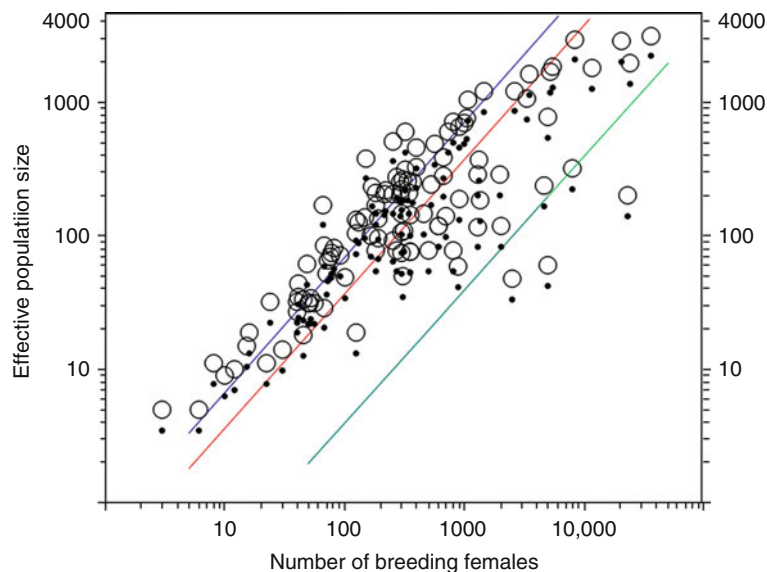
### Population Size and Structure

There are three ways to estimate effective population size ( $N_e$ ). Two make use of quantitative genetics theory [36].

First, from census counts of breeding males and females ( $N_m$ ,  $N_f$ ), typically through approximations developed for scenarios without selection, such as  $N_e \approx 4 \times (N_m \times N_f) / (N_m + N_f)$ , possibly expanded with information on variation in family structures. This method has been applied to Romanian, Japanese, and Croatian pig breeds [37–39]. The European Farm Animal Biodiversity Information System (EFABIS; <http://efabis.tzv.fal.de>) holds data on livestock breeds, including  $N_e$  derived this way. Figure 3 shows its estimated  $N_e$  values for 111 pig breeds, in relation to their  $N_f$  values. These datapoints are not on a straight line because of variation in the ( $N_f/N_m$ ) ratio; the reference lines indicate where  $N_f$  equals 5 (*top*), 10, or 100 (*bottom*) times  $N_m$ . Therefore, datapoints above the two highest reference lines represent situations with less than five (or ten) recorded breeding females per recorded breeding male – not very feasible scenarios in pig breeding, certainly at  $N_f$  values above 100. This only illustrates the difficulty of obtaining consistent census counts, and the limitations to deriving credible  $N_e$  estimates from them.

Second, through the rate of inbreeding ( $\Delta F$ ) as calculated from pedigree analysis, exploring the fact that  $N_e = 1/(2 \Delta F)$  under random mating. The main resource required here is a complete pedigree. This method has been applied to Japanese, American, German, and Finnish pig breeds [38, 40–42]. The approach can be taken an important step further by focusing on its across-breeds distribution characteristics [43].

A third approach makes use of molecular genetics technology, with various ways of analyzing marker data – for an overview see [44]. Álvarez et al. [45] analyzed the full pedigree of a small fragmented sheep population and genotyped microsatellites. They conclude that co-ancestry coefficients as estimated from pedigree data and from molecular data become rapidly correlated when pedigree depth increases, so that the expense of recording “molecular information in well established conservation programs may not be justified”; on the other hand, for small populations with a shallow pedigree “neither [pedigree] nor molecular information by themselves are sufficient [...]”; each available parameter offers partial information.”



**Pig Breeding for Increased Sustainability. Figure 3**

Reported numbers of breeding females ( $N_f$ ) in 111 European pig breeds, and the associated  $N_e$  values from  $4 \times (N_m \times N_f) / (N_m + N_f)$ ;  $N_m$  is the number of breeding males. The reference lines indicate where  $N_f$  equals 5 (*top*), 10, or 100 (*bottom*) times  $N_m$ . Open circles, raw data; small dots, multiplied by 0.7 in an attempt to adjust for the effects of selection (Data from EFABIS, December 2009)



For the correlation between pedigree-based and molecular measures of diversity “to be substantial, a considerable number of loci is required and, more importantly, a high variance of the [pedigree-based] inbreeding values should be present; [...] it should be preferable to use pedigree information whenever available, and limiting the use of markers to verify, correct, complete or even implement pedigree recording” [32].

Of course, in livestock breeds, deep and complete pedigrees are almost as scarce as dense DNA marker data, so “markers could be most useful in cases where little information on population history is available” [46]. Toro et al. [47] give a beautiful example of two LB varieties with complete 20-generation pedigree data, but these were maintained on an experimental farm – not a common situation.

Therefore, the molecular approach is widely seen as potentially more powerful than classical quantitative methods. Aspi et al. [48] write: “Owing to variation in family size and overlapping generations [...]  $N_e$  is [...] difficult to estimate from demographic field surveys. [...] Genetic methods may provide more effective ways for estimating  $N_e$ .” They genotyped microsatellites and estimated  $N_e$  at  $\sim 40$ . Moreover, from those same data “large genetic variation was found in the population despite a recent demographic bottleneck. No spatial population subdivision was found, even though a significant negative relationship between genetic relatedness and geographic distance suggested isolation by distance.” This is about a wolf population; similar quantification of  $N_e$ , demography, and subdivision would be very useful in livestock genetic resource management.

An effective population size of 40 would be regarded as dangerously low. Meuwissen [49] mentions a “critical effective size” (below which fitness steadily decreases) of 50–100, at least in populations not selected for traits negatively correlated to fitness. Ollivier et al. [50] transform  $N_e$  into the *extinction probability*  $P_{\text{ext}}$ , operationally defined as the expected level of inbreeding (accumulating deleterious mutations, which eventually leads to extinction) after 50 generations:  $P_{\text{ext}} = 1 - e^{-50/(2N_e)}$ . With a pig generation interval of 1 year, this goes back to [51] where risk of extinction is based on cumulative inbreeding over 50 years. The above  $N_e = 40$  works out as a dangerous 46% probability that the population will not survive

50 generations. A slightly different approach is the *degree of endangerment* described by [52]. The relationships among  $P_{\text{ext}}$ ,  $N_e$ , and  $\Delta F$  are described in more detail in [53–55], at different levels of complication. The latter source also considers the estimate’s accuracy, which will be useful once such estimates are to be used in practice.

Effective population size influences the extent of genetic drift, which affects the development of linkage disequilibrium (LD), i.e., the correlation between genotypes at different loci. Loci that are closer together are less often separated by recombination – so they are more strongly correlated, with higher LD values. In terms of data analysis of biallelic loci, the relevant relationships can be generalized [56] in terms of what is actually happening (due to  $N_e$ ) and of what can be observed (due to sample size) as

$$E(r^2) \approx \frac{1}{\alpha + kN_e} + \frac{1}{n} \quad (1)$$

$E(r^2)$  is the expectation of the square of the abovementioned correlation, a common parameter to quantify LD. Parameter  $\alpha$  equals 1 in the absence of mutation, and 2 if mutation is accounted for;  $k$  equals 4 for autosomes, and 2 for the X chromosome;  $c$  is the recombination rate between the markers being analyzed (their genetic distance from each other, in Morgan);  $n$  is sample size.

For an analysis of many autosomal markers, ignoring mutation, this works out as

$$N_e \approx \frac{1 - r^2}{4cr^2} + \frac{1}{4cr^2(nr^2 - 1)} \quad (2)$$

Hayes et al. [57] introduced *chromosome segment homozygosity* (CSH), an alternative parameter to  $r^2$  for quantifying LD. CSH has a smaller sampling variance than  $r^2$ , but the same approximate expectation (Eq. 1). When population size changes linearly over time, the effective population size of  $1/(2c)$  generations ago can be approximated as  $(1 - \text{CSH})/(4c \text{CSH})$ , the same form as Eq. 2 when  $n$  is large.

Accordingly, when many individuals are genotyped for many biallelic DNA markers varying widely in their mutual distance  $c$ , then the (supposedly linear) history of  $N_e$  can be traced by plotting  $\frac{1 - \text{CSH}}{4c \text{CSH}}$  against  $\frac{1}{2c}$ .

Amaral et al. [58] estimated LD decay (reduction of  $r^2$  with increasing  $c$ ) on SNP markers in ten European

and ten Chinese pig breeds (subsets of the breeds in Fig. 1). Figure 4 shows those data reworked into Hayes's relationship, using  $r^2$  instead of CSH, for the European breeds ( $N_e \geq 10,000$  for the Chinese breeds). Comparable data on commercial pig lines [59, 60, 226] has been added.

Figure 4 illustrates that this method can reveal interesting information about a population's demography. Such information would be most relevant to livestock genetic resource management when it covers recent generations, as [59, 60, 226] do. This requires large  $c$  values: from the equation above, the situation of two and five generations ago is represented by LD among markers 0.25 and 0.10 Morgan apart. The maximum distance between Amaral's markers was 0.03 Morgan.

This methodology is statistically demanding. England et al. [61] report on "simulations to show that [the most widely used LD] estimator is strongly biased when sample size is small [...] and below true  $N_e$ . This is probably due to [LD] generated

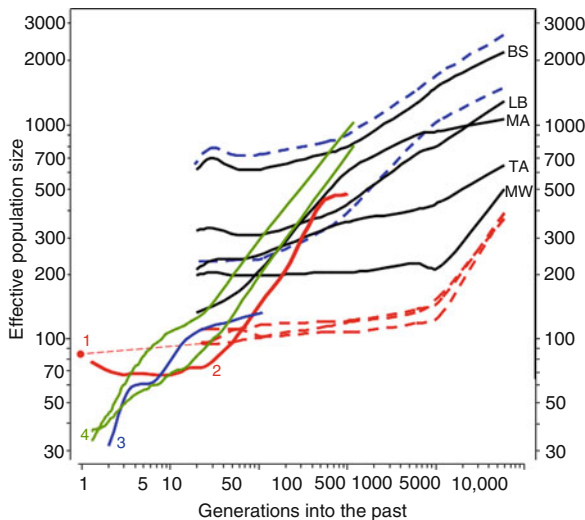
by the sampling process itself." They also proposed "a way to determine whether a given sample size exceeds population  $N_e$  and can therefore be used for computation of an unbiased estimate." Waples [62] confirmed this by describing how Eq. 1 above is inappropriate for low values of  $n$  and  $N_e$ , particularly when  $n < N_e$ .

The accuracy of the  $N_e$  estimator of Eq. 2 depends not only on the numbers of individuals and markers involved, but also on  $N_e$  and  $c$ . Based on [44], for a system with  $n$  individuals genotyped for  $m$  pairs of markers, each with a within-pair distance of  $c$  Morgan, an approximation of the standard error of estimated  $N_e$  is

$$\text{stderr}(N_e) \approx \left( N_e + \frac{2c(2-c)}{(1-c)^2 + c^2} \times \frac{N_e^2}{n} \right) \times \sqrt{\frac{2}{m}} \quad (3)$$

From this equation, estimates of  $N_e = 100$  or  $N_e = 200$  for five generations into the past ( $c = 0.1$  Morgan) would require  $n = 100$  individuals genotyped for  $m = 400$  or  $m = 750$  marker pairs to achieve a standard error of 10% of the estimate (i.e.,  $100 \pm 10$  or  $200 \pm 20$ ). Similar accuracies for one generation into the past ( $c = 0.5$  Morgan) would require  $m = 3,000$  or  $m = 10,000$  marker pairs.  $N_e$  estimates for more recent situations require more data for a given proportional accuracy. However, high accuracies (low standard errors) are easier achieved for lower  $N_e$  levels – often the more interesting ones from a resource management point of view.

Other molecular approaches to estimate  $N_e$  trace changes in allele frequencies over time (the *temporal* method), or derive (small) population size from (large) sampling errors of observed heterozygote proportions deviated from Hardy–Weinberg proportions (the *heterozygote excess* method). Schwartz et al. [63] give a review; further developments are described in [64–66]. The latter paper introduces the immigration rate into the equation – a useful issue in livestock scenarios with their common introgression of extraneous germplasm. Abdallah et al. [67] also take immigration into account when analyzing 12 of the breeds of Fig. 2. One of these was also analyzed by Amaral [58]; Fig. 4 gives Abdallah's estimate as the end point of the Hayes approach to Amaral's data.



**Pig Breeding for Increased Sustainability. Figure 4**

Time trends in effective population size for 14 European pig breeds, estimated from linkage disequilibrium decay patterns according to [57]. Blue dashed line, transboundary breeds; black solid line, local breeds (British Saddleback, Large Black, Tamworth, Middle White, Mangalica); Others, commercial lines: "1" from [67]; "2" from [59]; "3" from [60]; "4" from [226] (All other data from [58])

## Relative Utility

Comparisons of pig lines are difficult to organize and expensive to run: (semi)-governmental versions of such tests have run for several decades as *commercial product evaluation* in western Europe, focusing on growth and carcass traits. Reproductive traits are sometimes included by across-farm analysis of routine field data, with its inevitable and unpredictable bias. Robustness traits are very rarely included, not surprisingly given the demanding nature of categorical trait analysis. Such a test was designed [68] for growth and carcass traits to detect differences between genotypes of 0.25 trait standard deviations at 95% significance, with a statistical power of 75%; this led to a scheme with, on average, 67 sires per genotype, 2.6 litters per sire, and 1.8 tested pigs per litter. Subsequent tests also covered reproduction and longevity traits; these schemes specify 65 sires, 3 litters per sire, and 2.7 tested daughters per litter [69]. The statistical significance of breed differences must be tested against the variation among sires within breed, which requires inconvenient sampling schemes.

Gibson et al. [70] (condensed here) discuss characterization for production and robustness traits: this can only be genetically meaningful when the production environment is properly accounted for; environmental factors are so complex that records from different locations or times cannot be validly compared; valid breed comparisons are possible, first, when breeds are recorded simultaneously at the same location under identical management [as in the previous paragraph]; second, through meta-analyses linking records from different locations or times through overlapping breeds, statistically adjusting for environmental effects; such meta-analyses are powerful, but only valid when genotype $\times$ environment interactions are negligible (as in controlled confinement conditions); it will remain problematic that lifetime productivity traits are extremely difficult to record.

These authors stress that the functionality of information systems “must be greatly increased to allow extraction and customized analysis of phenotype and molecular genetic data within and between data sources; [...] breed information can be linked to [...] environment and production system mapping, allowing [...] disease resistance and adaptation traits

to be predicted from past and current breed distribution and use.” They conclude that “these are substantial but fully achievable functions.”

Tixier-Boichard et al. [46] emphasize characterization for robustness: “local breeds survive in harsh environments and this needs to be better understood; epidemics are major threats for all animal genetic resources across the world; climatic change is likely to increase the spread of tropical diseases to temperate areas. [Scientific evidence] that local breeds are adapted and resistant [...] has been obtained in several instances [...], well documented for parasitic diseases [...], with local breeds maintaining a better performance in the presence of parasites and/or exhibiting lower levels of parasite infestation [tolerance]”. These authors take a utilitarian position: “data on production systems, phenotypes and molecular markers should be used altogether in an integrated approach to characterization. [...] Decisions regarding conservation should incorporate all descriptors. Conserving without documenting would be useless.”

In summary, a livestock breed’s direct-use utility value should be assessed in terms of production traits and robustness traits like disease resistance/tolerance, adaptation to unfavorable conditions, and lifetime productivity. This assessment must use any available phenotypic and molecular data (the latter element comes back to the work of Ciobanu [33] and Iannuccelli [34], section “[Genetic Distinctiveness](#)”). All this will require extensive information systems, difficult to achieve but not impossible. Without such functionality, utility-directed breed conservation is not feasible.

Of course, phenotypic and molecular characterization can only focus on known characteristics. The main drawback is that the traits of true future interest (covering unexpected changes in consumer preferences, or in options for health or climate control or nutrition) cannot, by definition, be defined or measured. This makes it difficult to combine the utility issue with the insurance argument for conservation: it calls for option values. Viral diseases (e.g., PRRS, PMWS, influenza) are expected to break out on a wide scale every few years [71, 72], and any breed that happens to carry full resistance would suddenly have a very high utility value. This would require fast and widespread testing, and challenging logistics to distribute the relevant alleles throughout the worldwide pig industry – which

may well become feasible with further development of genomics and reproductive technologies. But that particular breed's utility is likely to drop dramatically again, as soon as the next outbreak (of a different virus) occurs.

### Urgency, Importance, and Feasibility: Priority Level of the Breed

To quote Wikipedia, Covey et al. [73] introduced “a framework for prioritizing work that is aimed at long-term goals, at the expense of tasks that appear to be urgent but are in fact less important.” This is about time management, but the same is relevant in global genetic resource management. Sections “Genetic Distinctiveness” to “Relative Utility” describe different features that might make a livestock breed a candidate for conservation – but limited funding requires prioritization. Obviously, highest conservation priority should be given to breeds that (1) have a great utility, (2) are strongly distinct from other breeds with much within-breed variation, and (3) are strongly endangered due to inadequate population size or structure. Items (1) and (2) are about *importance*; item (3) is about *urgency*. This requires integration of these issues, preferably quantitatively so that priority levels can be ranked and funding allocated. Another element to include is then (4) the cost of conservation – introducing the issue of *feasibility*: “identified benefits could be quantified so that society has some sense of how much the conservation is worth. Society can then determine how much they would want to spend on a conservation effort” [74].

A comprehensive way to deal with the above elements [75] defines the genetic distinctiveness of a breed (calculated with the relevant emphasis on between-breeds and within-breed diversity) as  $D$ , its utility value (for all relevant purposes) as  $U$ , recalls its extinction probability  $P_{\text{ext}}$  (based on its  $N_e$  and possibly on additional parameters) and defines the cost of reducing it (through any relevant conservation action) by  $\Delta P_{\text{ext}}$  units as  $C$ . Then the priority ranking  $R$  for such a conservation action would simply be

$$R = \frac{(D + U) \times \Delta P_{\text{ext}}}{C} \quad (4)$$

This would be calculated for every breed in the conservation portfolio, assuming that all their  $D$ ,  $U$ ,

$P_{\text{ext}}$ , and  $C$  values can be directly compared – which requires  $D$  and  $U$  to be expressed in the same unit, and everything to be calculated using the same algorithm and parameter definitions across breeds. Breeds with higher  $R$  values get a higher priority for conservation – because their conservation is more important ( $D$ ,  $U$ ), more urgent ( $P_{\text{ext}}$ ), and/or more feasible ( $C$ ). The probability  $P_{\text{ext}}$  takes values from 0 to 1, and  $\Delta P_{\text{ext}}$  from 0 to  $P_{\text{ext}}$ : reducing the extinction probability by its full value (i.e.,  $\Delta P_{\text{ext}} = P_{\text{ext}}$ ) comes down to safeguarding the breed entirely.

These authors notice that the costs of the most complicated in vivo, in situ conservation schemes would likely be proportional to conservation effort, i.e.,  $\Delta P_{\text{ext}}/C$  is roughly constant and ranking is based on  $D$  and  $U$ . By contrast, the costs of the simplest in vitro cryoconservation schemes might vary only little, i.e.,  $C$  is roughly constant and ranking for complete safeguarding is based on the *cryoconservation potential*  $(D+U) \times P_{\text{ext}}$ . Real-life conservation programs would fall between these extremes.

Simianer et al. [76] compared three forms of the actual relationship between  $C$  and  $\Delta P_{\text{ext}}$ , applied these to a set of breeds characterized in terms of genetic distances and extinction probabilities, and found that “conservation funds should be spent on only three to nine of the 23 breeds, depending on the model used.” This approach was further formalized [77] into a comparison of *maximum-risk*, *maximum-diversity*, and *maximum-utility* strategies to determine the optimum set of breeds to conserve, which favors the latter strategy which combines diversity and utility – although it obviously requires the quantification of  $U$ , which makes it difficult to implement in practice. The latter two approaches quantify the *expected conserved diversity* or the *expected conserved utility* of possible sets of breeds that may be successfully conserved at some point in the future, and then calculate the *marginal diversity* or *marginal utility* of each breed by differentiating with respect to the breed's  $P_{\text{ext}}$ . These are then multiplied by  $P_{\text{ext}}$  to obtain the breed's conservation potential, similar to the term  $(D+U) \times P_{\text{ext}}$  of Eq. 4.

As argued at the end of the section on “Relative Utility,” utility is the weakest element here: conservation aims at the future, and future utility cannot be predicted. All the other elements can in principle be dealt with by a complete pedigree and/or dense DNA

samples from ~100 individuals. One of the options would be to drop U from the equation and rank conservation priorities on D only (the *maximum-diversity* option of above). This reasoning is taken to its logical extreme by suggesting to “devote the majority of present conservation budgets to freezing [...] samples from existing breeds [...] concentrate on ova and sperm from abattoir material, and somatic cells, e.g., ear clips (the latter in anticipation of the increasing effectiveness of somatic cloning),” because future “genomic tools will open up completely novel means of exploiting genetic resources” [14]. In line with this, the USA has “invested in the establishment of an in vitro conservation program and a genebank [covering 18 local pig breeds and one ITB at the time of reporting]. Collections are being built up very quickly, in close collaboration with the industry. Breeding companies use the genebank as a backup of their breeding work. In Canada, a program for in vitro conservation [...] will be implemented in the near future” [5], as later documented by [www.ushrl.saa.ars.usda.gov/SP2UserFiles/Place/54020500/documents/update%208-10-02.pdf](http://www.ushrl.saa.ars.usda.gov/SP2UserFiles/Place/54020500/documents/update%208-10-02.pdf) and [dsp-psd.pwgsc.gc.ca/collection\\_2008/agr/A52-88-2008E.pdf](http://dsp-psd.pwgsc.gc.ca/collection_2008/agr/A52-88-2008E.pdf).

### Target Traits for Genetic Improvement

Livestock breed improvement requires a breeding goal and a selection strategy best suited to the needs of the production system and the market that it supplies – this is one of the elements of *population-level genetic resource management*. Animal breeding technology has a long tradition in this field, but improvement of an endangered breed may require substantially different goals and strategies than in mainstream industry breeding programs. FAO [5] write: “the most appropriate strategies for managing these breeds may involve only limited genetic change [...] to maintain adaptation to the local environment and disease challenges, and [...] to maintain the level of a production trait [...] if this is currently [near] an optimum level.” Of course, the insurance argument for breed conservation assumes that the breed may, at some time, support a different production system than its original one, because it happens to carry alleles (likely adaptive ones) of large utility, then and there. This adaptive utility must be balanced with production utility,

which will likely be much lower than in commercial lines. The actual uptake of the breed as a source of adaptive alleles will be much easier when the production-related lag is limited. So for the insurance argument, genetic improvement of production traits is desirable – but not at the expense of adaptive quality: possible genetic antagonisms between production traits and anything else require specific attention. “The genetic basis of population differentiation for fitness traits will be non-additive, with different adaptive gene complexes evolved in each breed. Genetic improvement programs therefore should start with an adapted population, with selection then for production traits” [78].

Another argument in favor of improvement of production traits is that many endangered breeds are endangered precisely because they lag behind other breeds in terms of production traits. The common idea that improvement of production traits inevitably leads to a reduction of robustness is false – it just requires a sensible breeding program, see the section on “Robustness.”

### Genetic Improvement Programs

As argued above, any livestock breed with a future must support an agricultural production system. Rege [79] writes: “the most rational and sustainable way to conserve animal genetic resources is to ensure that indigenous breeds remain functional parts of production systems, that is, *conservation through use*. This is possible only if economically important attributes of indigenous breeds are identified, studied and incorporated in breed improvement programmes.” Genetic improvement will always be required – unless the production system is completely static, without any interactions with the world around it. A useful concrete example is from [80]: “during the past few years the Limousin pig, endangered and neglected during the 1970s and 1980s in favor of the better-performing large breeds, has become popular with consumers looking for quality. Today it is victim to its own success with supply being lower than demand, so that it has become necessary to develop its productivity” (translated).

This requires a breeding goal and a selection strategy best suited to the needs of the production system and the market that it supplies, as discussed in the

Section on “[Target Traits for Genetic Improvement](#).” A checklist of items that play a role here, from definition of the product and the market to evaluation of the breeding program’s profitability, is in [81].

The breed also has to be *maintained* – preserving as much of its genetic variation as is feasible. And its qualities must be *exploited* in an optimal way, making efficient use of other resources.

Livestock breed *maintenance* requires strategies and tools to keep  $N_e$  sufficiently high. This comes down to keeping inbreeding under control, usually by minimizing co-ancestry in the breeding population. There is a possible antagonism with the previous point: selection in a population reduces  $N_e$ , so a balance will have to be found and maintained. There are many rules-of-thumb to delay inbreeding (e.g., *keep one son from every sire* or *maximize generation intervals*); an example of implementation in commercial lines is in [82]. Many of these rules perform well on the short term but have unexpected long-term effects, and they usually reduce genetic improvement unpredictably, leading to uncertain genetic resource management.  $N_e$  can be affected by age at first breeding and culling policy in quite counterintuitive ways [83], so that “the general tendency is contrary to the expectation that [ $N_e$ ] would increase with increasing [longevity].” High longevity increases generation length but reduces genetic drift; the combined effect favors a short productive lifetime, so that  $N_e$  can actually be increased by early culling.

A better solution is to apply co-ancestry management and mate selection based on optimum contribution theory or similar frameworks. The principles are covered in detail in [49, 84, 85]. The latter source’s *mate selection index* (MSI) is an optimized criterion, based on an objective function such as [86]:

$$\text{MSI} = \text{EBV} - \lambda_1 x' Ax - \lambda_2 \bar{F} \quad (5)$$

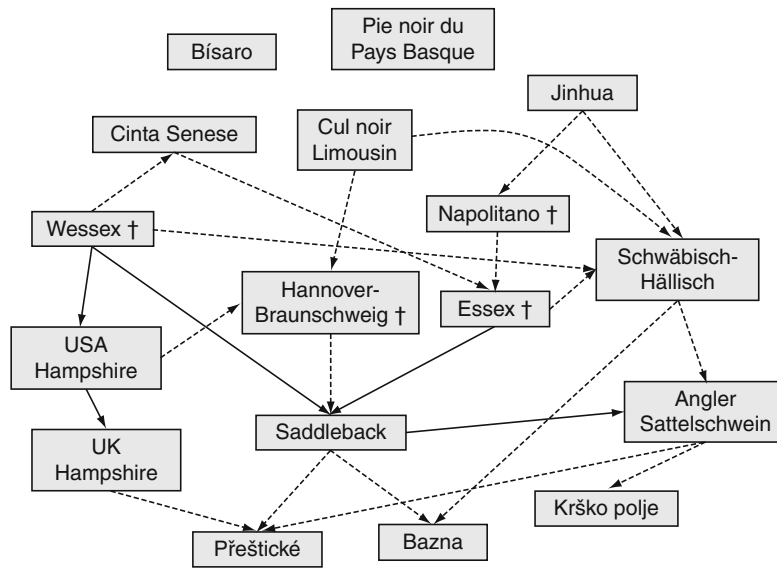
Here, (1) EBV holds estimated breeding values; (2)  $x' Ax$  represents average co-ancestry in the system;  $A$  holds the additive relationships among all animals, weighted by their contributions  $x$  (numbers of progeny) to the next generation; (3)  $\bar{F}$  is their average inbreeding coefficient, calculated from  $A$ ;  $\lambda_1$  and  $\lambda_2$  are positive weighting factors. This is a cost-benefit equation, with element (1) representing the benefit and (2) and (3) representing the (genetic) cost: EBV,

$A$ , and  $\bar{F}$  are known, so the system can be solved to deliver the contributions  $x$  that give the best (genetic) cost-benefit for particular values of  $\lambda_1$  and  $\lambda_2$ . The result of this is a list, based on  $x$ , with animals to select and animals to mate to each other (the optimum selections and optimum matings that lead to optimum contributions).

Proper genetic resource management can lift populations much smaller than the FAO threshold of 1,000 breeding females to the “not at risk” level. Many industrial pig lines are maintained at far less than that population size (e.g., see Fig. 4) with a secure genetic future. Careful immigration and admixture is regular practice in the breeding industry, but it is often avoided by breed societies and similar structures for chauvinistic reasons.

For example, Fig. 5 shows an intensive network of historical genetic connections among the European black-belted and saddled pig breeds, with a genetically useful loop with Hampshire (an ITB). Many of these populations have a herd size far lower than 1,000 sows, and most of those focus on the same set of characteristics (meat quality and robustness, and of course coat color) when describing their distinctiveness in the associated Web sites. This would suggest a clear case for the quantification of the various elements of Eq. 4 above, followed by regular and careful genetic exchange to make resource management easier.

*Exploitation:* A common negative term in the literature on genetic resource management is *indiscriminate crossbreeding*. For example, “very often, crossbreeding has been indiscriminate and the local breeds that underpin the crossbreeding program have been lost because of a lack of understanding [...] that these pure breeds must be maintained to support the system” [87]. This is clearly a management issue. On the other hand, these authors mention systems where crossbreeding is “used for gradual breed replacement with [...] the controlled [...] formation of composites [...] for specific production systems,” as “a rapid method of introducing desirable traits into local well-adapted breeds,” and “as a way out of a narrowed genetic base in commercial breeds.” The logical way of making use of exotic germplasm without any risk of endangering the LBs is “structured cross-breeding systems, such as ‘terminal crossing’ where [ $F_1$ ] animals are slaughtered or where specialized crossbred dam lines are used.”



**Pig Breeding for Increased Sustainability. Figure 5**

Genetic connections from ~1,770 (Jinhua) to present, among black-belted and saddled pig breeds. Consecutive series of arrows do not imply any chronological development. Solid arrows represent the main source of a breed; dashed arrows imply the existence of other sources not included here. (Data from [16, 37, 222–225]; Lenoir H, 2010, IFIP Institute du Porc, Le Rheu, France, Personal communication; [www.elbarn.org](http://www.elbarn.org); [www.besh.de](http://www.besh.de); [efabis.tzv.fal.de](http://efabis.tzv.fal.de); [dad-training.fao.org/cgi-bin/EfabisWeb.cgi](http://dad-training.fao.org/cgi-bin/EfabisWeb.cgi); [www.thepigsite.com/info/swinebreeds.php](http://www.thepigsite.com/info/swinebreeds.php))

Pig examples are from Germany where Angler Sattelschwein, Schwäbisch-Hällisch, and Bunte Bentheimer are crossed with Pietrain; and from Spain where various Iberico strains are crossed with Duroc, all to optimize meat quality versus meat quantity in the terminal F<sub>1</sub> product.

This chapter is about science and technology. But “logistics, not science, is the underpinning of a successful breeding policy. Without a system for handling the details of livestock identification, classification and movement, the science is of little avail” [88]. This certainly holds for endangered populations, typically managed by fragmented groups of independent-thinking people. Nimbkar et al. [87] stress the importance of “structures to organize the keepers of animals and help motivate communal efforts [...] allowing livestock keepers better access to information, [...] extension services, facilitating the organization of training, and improving [...] marketing. In Europe, there are strong farmer cooperatives and breeding organizations that go back a century.” Indeed, successful breeding programs were founded on equally fragmented conditions in many European countries

and also in Canada, starting out with breeds with similar production performance as the ones from that same area that are now endangered or extinct. Apart from the clear need for incentive and for institutional backing, there are two prerequisites for such development. First, technically: an efficient system for ensuring genetic connections among farms, e.g., through regular exchange of males or across-farm AI – everything else of a technical nature will have to build upon that. Second, organizationally [89]: employment of professional genetic expertise by the breeders’ organization – so that the system does not have to rely on fragmented and unstable governmental service, and can arrange for effective feedback between breeders’ objectives and technological options.

## Pollution

FAO’s *Livestock’s long shadow: environmental issues and options* report [90] gives an overview of the amount of nitrous oxide (N<sub>2</sub>O) released from livestock manure and urine, worldwide, in 2004. Of the total emission of  $3.69 \times 10^9$  kg, 12% was due to pigs – just over half of

that from Asia.  $N_2O$  is an effective greenhouse gas (GHG), also involved in the depletion of the ozone layer. Other nitrogen compounds that enter the environment from livestock excreta are ammonia ( $NH_3$ ) and nitrogen oxides ( $NO$  and  $NO_2$ ), involved in acidification or, indirectly, in global warming.

### Technology

The FAO report [90] devotes much text to options for reduction of nitrogen emission, most of which involve manure management and improved animal nutrition. For example (p. 122): “An important mitigation pathway lies in raising the low animal nitrogen assimilation efficiency [...] through more balanced feeding (i.e., by optimizing proteins or amino acids to match the exact requirements of individual animals or animal groups). Improved feeding practices also include [...] improving the feed conversion ratio [FCR] by tailoring feed to physiological requirements. However, even when good management practices are used to minimize nitrogen excretion, large quantities still remain in the manure.” This is quantified in another table in the report (p. 137) with typical values for nitrogen intake, retention, and excretion in cattle, pigs, and poultry. According to these numbers (which go back to [91]), across these species, only about 19% of nitrogen ingested in “less productive situations” is retained in meat, eggs, and/or milk – in “highly productive situations,” this goes up to 30%. Likewise nitrogen retention rates ( $N_{ret}$ ) of about 34% for pigs were reported for the 1995 “highly productive situations” of France, Denmark, and the Netherlands [92].

These retention rates are indeed low, but the difference between the above productivity levels is considerable, suggesting scope for increase by “improved feeding practices”. Dourmad et al. [93] state that “the ultimate reduction of N excretion can be reached when multi-phase feeding is combined with a perfect balance of essential amino acids and [...] optimization of the supply of non-essential amino acids” – ideally on a daily basis. They refer to a 1995 experiment [94] where the use of a single diet over a growing period from 26 to 101 kg body weight was compared to such an optimized multiphase feeding strategy, and where  $N_{ret}$  values at 34% (single diet) and 50% (optimized regime) of the ingested nitrogen were found.

Later studies in laboratory conditions have achieved higher retention rates. De Lange et al. [95] studied the effect of dietary amino acid levels on protein deposition rate in pigs from 39 to 77 kg liveweight, and present results that lead to  $N_{ret} = 61\%$  at complete amino acid availability – “at a more typical protein digestibility, this would become 56%” (cf. de Lange, personal communication, 2010). Similarly, Buraczewska et al. [96] measured  $N_{ret}$  of up to 57% in 35-kg and 45-kg pigs of a “high lean gain potential” genotype, after optimization of the dietary amino acid composition. In more practical conditions, Pomar et al. [97] fed pigs from 25 to 105 kg liveweight according to a “traditional three-phase feeding program” or with “individually tailored diets,” obtaining  $N_{ret} = 37\%$  and 48%, respectively.

Curiously, the FAO report [90] pays no attention to a logical alternative to “improving [FCR] by tailoring feed to physiological requirements,” i.e., improving it by tailoring these physiological requirements themselves, through animal breeding [98]. When nitrogen excretion data of growing (Landrace  $\times$  Large White) and (Hampshire  $\times$  Duroc) pigs were adjusted for body weight and feed intake, significantly different values were obtained between the genotypes – which reveals genetic variation so that “genetic selection may be an effective method for altering nutrient utilization and output” [98]. Heritabilities for laying-hen excretion traits such as dry excreta weight, excreta humidity rate, and the ratios of dry excreta and nitrogen excreta to feed intake (0.25–0.46) and for dairy cow methane production (0.12) were reported in [99, 100].

Improvement of traits like litter size, sow feed intake, growth rate, and FCR reduces nitrogen excretion ( $N_{excr}$ ) from sows and growing pigs [101]. The 1988–2007 genetic trends for growth rate (+8.5 g/day/year), FCR (–0.02 kg/kg/year), and litter size (0.16 pigs/litter/year) in the UK pig sector were estimated to cause 0.8% annual reduction of the associated global warming potential of GHG emission [102]. In that study, the genetic reduction of FCR explains about 70% of the reduction in  $N_2O$  emission; the genetic increase of growth rate explains about 70% of the reduction in  $NH_3$  and methane emission. The genetic increase of litter size (which reduces the sow herd with its emission, for a fixed number of slaughter pigs) explains about 20% of all three elements. The future

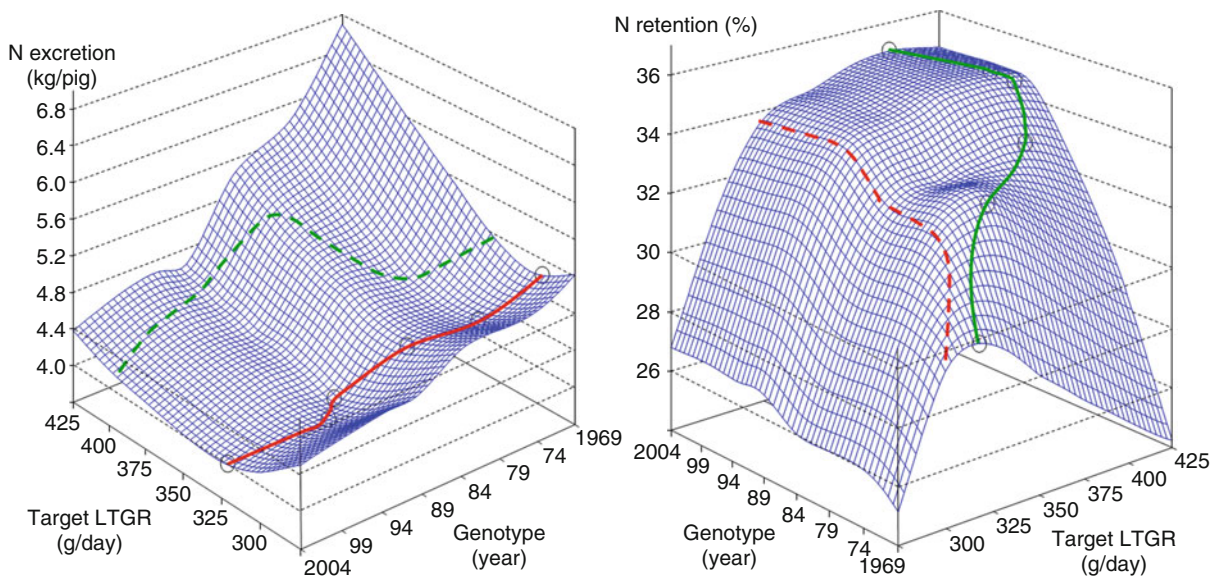


scope for emission reduction is underestimated in that study, for two reasons. First: genetic trend of lean content (which was substantial in the UK during this period, probably about 0.5%/year) was not taken into account. Second: future trends in all these traits may be expected to be stronger than the historic values, due to further development of genetic technology and of the “uptake rate of improved genetics by the commercial level” [102].

The reduction of  $N_{\text{excr}}$  due to genetic improvement of production traits is quantified by the simulation results in Fig. 6, obtained with the model of [103]. The time trends of model parameters such as maximum protein deposition rate ( $PD_{\text{max}}$ ) that were described for six pig sire lines [104] can be used to model  $N_{\text{ret}}$  and  $N_{\text{excr}}$  at the nucleus level throughout the 1969–2004 period. The simulations involve these six progressively advanced genotypes, grown from 20

to 120 kg body weight. Each of these genotypes was fed ad libitum on each of seven three-phase (20–50, 50–80, and 80–120 kg) diet specifications, targeting overall lean tissue growth rates (LTGR) from 275 to 425 g/day in steps of 25 g/day [105]. This involves diets with a fixed digestible energy content ( $DE = 14.2 \text{ MJ/kg}$ ) and varying levels of crude protein (e.g., from 12.2% to 15.5% in phase 3) and essential amino acids (e.g., lysine from 0.525 to 0.765% in the diet, ditto).

The older genotypes in this simulation do not have the potential to achieve the higher LTGR targets, with the consequence of low  $N_{\text{ret}}$  and high  $N_{\text{excr}}$  levels. Low LTGR targets obviously lead to low  $N_{\text{ret}}$  as well, more so in the older genotypes. Figure 6 shows clear optimum trajectories across genotypes and feeding strategies, for both  $N_{\text{ret}}$  and  $N_{\text{excr}}$ . It also shows that these optimum trajectories follow different paths for both characteristics, particularly in the more advanced genotypes.



**Pig Breeding for Increased Sustainability. Figure 6**

Nitrogen excretion (*left*) and retention rate (*right*) in simulated growing pigs of six genotypes (representing sire lines from 1969, 1976, 1984, 1990, 1993, and 2004; From [104]). Each genotype was fed from 20 to 120 kg liveweight on three-phase diet specifications targeting [275, 300, ..., 400, 425] g/day lean tissue growth rate (LTGR), according to [105] – the higher LTGR targets are beyond the potential of the older genotypes, and the more advanced genotypes do not realize their potential due to inadequate nutrient supply. The blue response surfaces are spline interpolation plots through the  $6 \times 7 = 42$  simulated datapoints. The red and green trend lines represent minimum excretion and maximum retention, respectively, for each genotype. Each broken trend line mirrors the solid one of the same color (with its datapoints as black circles) in the other graph, approximately

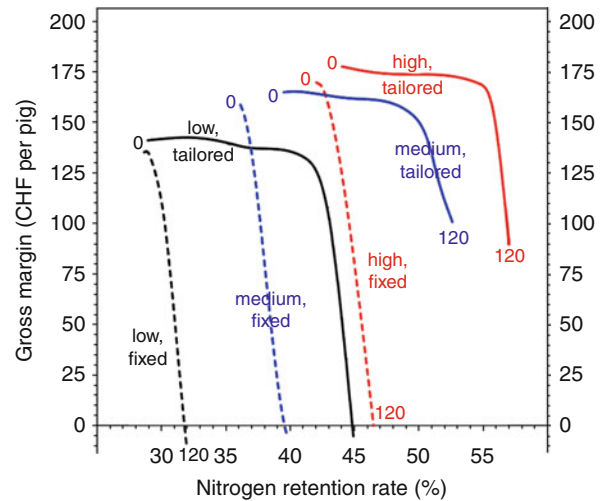
Along these optimum trajectories (i.e., when fed to achieve maximum  $N_{ret}$ , or minimum  $N_{excr}$  within the limits of the three-phase feeding program) the 2004 genotype shows a proportionally 19% higher  $N_{ret}$  or, alternatively, a 20% lower  $N_{excr}$  than the 1969 genotype. Deviations from these optimum diet composition settings have much stronger effects in the older genotypes than in the more advanced ones.

The more advanced genotypes in this simulation show a clear diminishing-returns pattern for  $N_{ret}$  with a very flat asymptote around  $N_{ret} = 35\%$ . But the simulated  $PD_{max}$  levels (i.e., genetic potential for  $N_{ret}$ ) increase progressively throughout the 1969–2004 period covered here – so the expression of that potential must be constrained by the three-phase fixed-diet program employed here. This comes back to the real-life results of Pomar et al. ([97]; above) who obtained  $N_{ret} = 37\%$  with a three-phase fixed-diet program and  $N_{ret} = 48\%$  with “individually tailored diets.” The latter strategy involved a daily analysis of the performance-to-date of each pig, to predict today’s individual body weight, growth rate, nutrient requirements, and ad libitum feed intake. Each pig was then fed a mixture of basic rations via automated feeders to match these predictions.

It follows that the more advanced genotypes require more advanced feeding strategies to bring their more sustainable performance potential to expression.

This was further explored by Morel and Wood [106], who used simulation modeling to quantify nitrogen flux in growing pigs of low, medium, and high LTGR genotypes, on a variety of dietary regimes. Their simulated high LTGR genotype achieved  $N_{ret} = 57\%$  when fed daily individually tailored diets (similar to Pomar et al.’s [97] diets mentioned above) with a strong focus on the minimization of  $N_{excr}$ . By contrast, the simulated low LTGR genotype achieved  $N_{ret} = 29\%$  on a three-phase fixed-diet program with all focus on gross margin (i.e., carcass return minus feed costs). So the various  $N_{ret}$  results of these simulations (see Fig. 7) span the relevant range of commercial and high-tech conditions described above.

These authors conclude that although “a reduction in nitrogen excretion is mainly achieved through a reduction in nitrogen intake” (as in Fig. 6), “genotypes with a high lean growth potential can be more profitable [in terms of gross margin] and have



**Pig Breeding for Increased Sustainability. Figure 7** Gross margin (from growth rate, backfat depth, and feed intake) in relation to nitrogen retention rate ( $N_{ret}$ ) in simulated growing pigs of three genotypes (low, medium, and high lean tissue growth rate), fed three-phase fixed diets or individually tailored rations. The labels “0” and “120” indicate the extremes of a range of weighting [ $N_{ret}$ :margin] from [0:1] to [120:1] in the objective function for diet optimization. CHF: Swiss Francs (Data from [106] and Morel PCH, 2010, personal communication)

improved nitrogen excretion.” A simple ANOVA of their results of gross margin and  $N_{ret}$  across the simulated scenarios (supplementary data from Morel PCH, 2010, personal communication) quantifies this. The low versus high LTGR genotypes show least-square means for simulated gross margin at 102.0 versus 150.0 CHF per pig, and for  $N_{ret}$  at 34.1% versus 46.8%, respectively. Their “three-phase fixed-diet” versus “individually tailored” feeding regimes show least-square means for gross margin at 116.3 versus 138.6 CHF, and for  $N_{ret}$  at 36.7% versus 45.0%, respectively. So Morel’s genetic scenarios [106] are *more* effective (because further apart) than his feeding regimes, for improving both gross margin and  $N_{ret}$ . With some generalization, this can be put into perspective as follows.

With the 1969–2004 trend of  $PD_{max}$  in pig sire lines from [104] (above), the  $PD_{max}$  input values of Morel’s simulated genotypes (120, 160, and 200 g/day) can be

located in time at 1970, 1987, and 1997, respectively – so his low and high LTGR genotypes are 27 years of sire line genetic improvement apart. The simulated gross margin of those genotypes (three-phase fixed-diet, full focus on gross margin) differs by 35.7 CHF (i.e., about 23 EUR).

Time trends of growth and carcass traits in commercially available slaughter pig genotypes [107] can be converted to gross margin trends as in Fig. 8. The average range for a contemporary comparison of seven genotypes (the most common configuration in this data, and also a fair representation of practical local availability of genotypes) is 10 EUR per pig.

This provides an (under)estimate of the range in gross margin among the various slaughter pig genotypes that are commercially available at any point in time (an underestimate, because these CPE trials do not include *all* available genotypes, particularly not the less advanced ones). This number is equivalent to  $10/23 = 43\%$  of the difference between Morel's [106] low and high LTGR genotypes. Combined with the above surmise that “Morel's genetic scenarios are more effective [...] than his feeding regimes are, both

for improving gross margin and for improving  $N_{ret}$ ,” it follows that a well-informed choice of the most appropriate slaughter pig genotype *available at any point in time* will have almost half of the influence on  $N_{ret}$  (and on margin) that full implementation of individually tailored diet optimization will have.

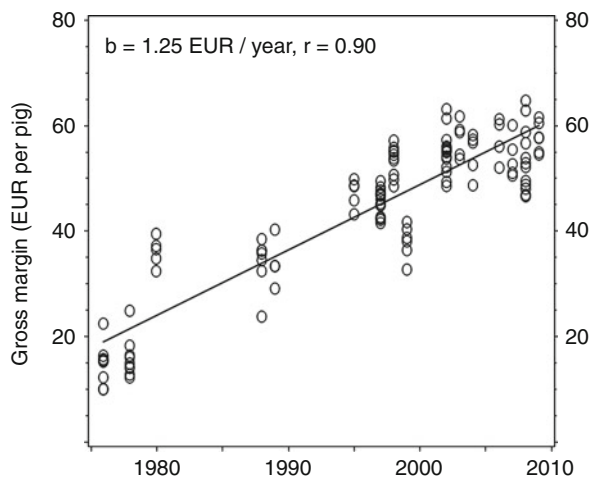
### Strategies

The above section shows that pig breeding in general makes a considerable contribution to the reduction of nitrogen emission from growing pigs. Up to now, these effects have not been incorporated into formal breeding goals, so they are not under control and cannot be credited, as such. Many governments have an increasingly active policy of pollution reduction, and the global warming potential of the excreta from livestock production is under increasing scrutiny. This has led to a series of economic studies into the effectiveness of taxation of nitrogen (and other chemicals) emission on the farm level, focusing on countries as politically (and productively) different as Italy, Switzerland, and the Netherlands [108–110].

A central element in the associated econometric approaches is the *shadow price* of nitrogen emission. Paul et al. [111] describe the “shadow values for the bad outputs” of an agricultural production system as “the marginal amount that producers [...] would be willing to pay for unrestricted use of the environment” to dispose of those bad outputs – e.g., for the right to increase their bad outputs in a situation where legislation attempts to reduce these.

Operationally, shadow prices are estimated as the partial derivative of the producer's profit equation with respect to the bad output factor – equivalent to marginal economic values for production traits in animal breeding. These are demanding statistics. Key et al. [112] studied the productivity of the USA slaughter pig sector, and note that estimating the “shadow price of manure [...] with the data available would require making a set of assumptions that would likely introduce substantial error [...], an accounting of hog farm output that includes manure is left for future research.”

The shadow price of nitrogen emission from growing pigs, sows, and dairy cows in the Netherlands was estimated at 2.7, 10.8, and 6.5 NLG/kg, respectively [110, 113].



**Pig Breeding for Increased Sustainability. Figure 8** Gross margin (from growth rate, estimated lean content, and feed intake) in slaughter pigs of commercially available genotypes, as recorded 1976–2009 in public Commercial Product Evaluation trials. The trend line is the linear regression. EUR: European Euros (Data from [107])

The UK government has included the shadow price of GHGs as a structural element of its cost-benefit evaluation of any policy that it funds or supports, using values based on [114] where a shadow price of CO<sub>2</sub> is derived that can be converted to a shadow price of nitrogen (as a component of N<sub>2</sub>O) at 9.25 GBP/kg.

Obviously, such shadow prices depend on practically every feature of the production system and its surrounding conditions. Using again Paul et al.'s [111] description above, the amount that producers would be willing to pay for the right to increase their bad outputs in a situation where legislation attempts to reduce these, would depend on (1) how strongly this legislation attempts to enforce the reduction, and (2) the impact of such a reduction on the remaining elements of the producer's profitability. Point (1) is largely a political factor; the UK government was criticized by environmental NGOs for the supposedly artificially low value of its abovementioned shadow price of CO<sub>2</sub>. Point (2) may well lead the producer to decide to *not* reduce his bad outputs, because that is still more profitable – as was the case in Paul's study which focused on pesticide usage: her shadow price estimates were negative.

Wall et al. [115] notice the equivalence of the shadow price of a bad output and the marginal economic value of a production trait, in the sense that both can be used to weight their characteristic into a breeding goal. They refer to the EU-ETS Emissions Trading Scheme that was set up to support the EU to meet its commitments to the Kyoto Protocol ([ec.europa.eu/environment/climat/emission/index\\_en.htm](http://ec.europa.eu/environment/climat/emission/index_en.htm)), which will effectively determine the shadow price of GHG emission in the EU. They write “suppose [...] that agriculture is forced into an [ETS] and that farmers must hold valuable permits either through initial allocation or by purchasing in the ETS. [This] will immediately move GHG mitigation traits from a public to a private breeding objective [...]; the prevailing emissions price becomes the relevant economic weight that should be incorporated in any breeding index that includes mitigation potential.” The demand for “breeding indexes that include mitigation potential” will become very concrete, once pig producers have to deal with such a scheme – in line with this, the total costs of 2007 environmental government policy in the Netherlands were estimated at

“around 0.11 EUR per kg [carcass] weight, of which 0.08 EUR was for manure disposal. In 2013, these costs will be 0.02 EUR higher as a result of the ammonia emission reduction policy” [116].

Figures 6 and 7 show that such a pig breeding policy is technically feasible: Nitrogen retention rate is favorably correlated with the conventional production traits and can easily be included into breeding goals and selection strategies.

## Animal Welfare

In 1976, the member states of the European Community ratified the *European convention for the protection of animals kept for farming purposes*, regulating that livestock must be properly housed, fed, and cared for. In 1992, the following text was added (condensed here for clarity): “Breeding procedures which may cause suffering or injury to animals shall not be practiced. No animal shall be kept unless it can be expected, on the basis of its phenotype or genotype, that it can be kept without detrimental effects on its health or welfare.” Five years later again, the EC's Scientific Veterinary Committee recommended that “no selection should occur without reference to the effects of that selection on welfare of [pigs]. The continuation of new genetic lines in which the welfare of the animals is, on average, worse than that of existing lines should not be permitted” [117].

Such statements leave the impression that animal breeding may be bad for the animals involved. This goes back to the late 1970s when animal breeding technology became much more powerful than before, due to improved data recording and processing (BLUP and, above all, computing power), and improved reproductive technology. Simultaneously, animal production in the western world experienced strong intensification connected to a long period of low, volatile, and unpredictable farm profitability [118]. The production sector therefore developed a strong and focused demand for animals with improved production performance.

This led to a strong and effective focus on production traits in livestock breeding in that period, with less attention for animal robustness traits. This lack of balance has caused fitness constraints in pigs, particularly in environmental conditions inadequate to

support the improved production potential, and particularly before this problem was being dealt with in modern pig breeding.

Intensification of animal production has also led to housing and management conditions that shelter the pig from climatic, nutritional, parasitic, and predatory challenges – but compromise much of the expression of its instinctive behavioral repertoire. This deprivation leads to frustration, with welfare problems for the affected individual and for its penmates.

Fitness constraints and deprivation are among the “criteria for determining the ethical limit for genetic selection” [119]. A third issue is formed by the routine invasive treatments that have been common, worldwide, since the onset of animal domestication – in pigs mainly tail docking and castration. These aim at a pragmatic (but unrefined and reoccurring) reduction of undesirable features that can also be dealt with through animal breeding – more complicated, but permanent and less physically invasive.

While worldwide demand for pig meat increases, it will become more and more relevant to resolve problems with pig welfare in intensive production systems. It is unlikely that this increasing demand will be met by any other production system than intensive ones, particularly in Latin America, Russia, and Asia. In accordance with FAWC’s [120] plea for “a greater emphasis in breeding programs on traits associated with good welfare,” animal breeding and genetics technology can then contribute in three areas: (1) robustness, (2) deprivation, and (3) avoidance of invasive treatments. In practice, each of these areas raises (4) ethical arguments, sometimes intense enough to dominate the issue – so although this chapter is about technology, these will be addressed as well. Sections “[Robustness](#)” to “[Ethical Aspects](#)” deal with these areas in this order.

## Robustness

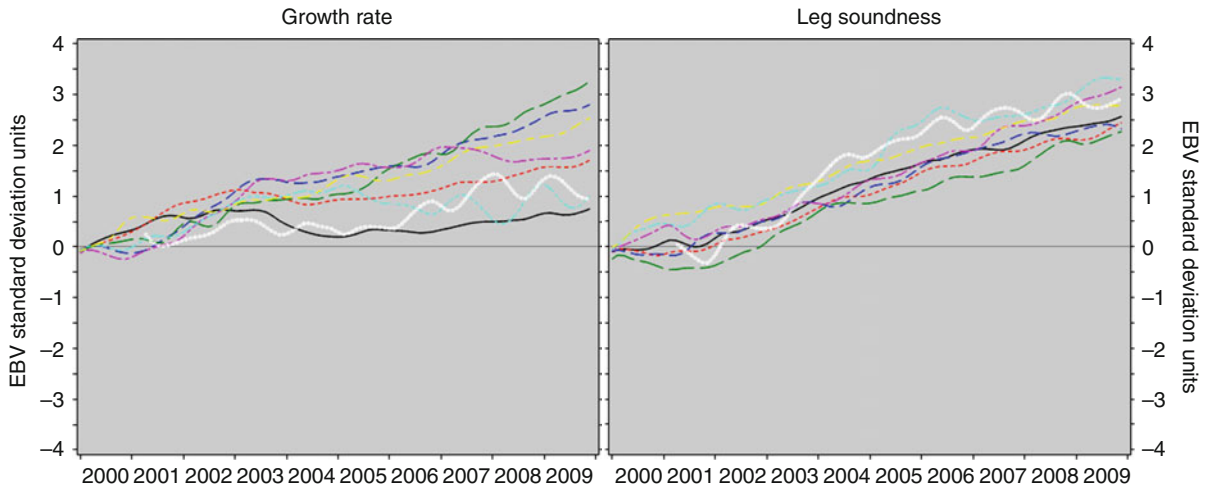
When animals of high-performance genotypes are kept in production systems that are inadequate to provide the resources they need to express their potential, the animals commonly show disturbed resource allocation [121] and functional disorders of the skeletal and cardiovascular systems, muscle physiology, the reproductive system, or immunocompetence [122]. For pigs,

obvious indicators of reduced animal welfare in this respect are (1) increased mortality rates and reduced sow longevity, (2) disease (morbidity), and (3) lameness.

The issue here is one of environmental sensitivity. There are two ways to deal with the problem: make the environment more resource providing or make the genotype less sensitive. There are two strategies for the latter: (1) direct selection for robustness traits and (2) selection against environmental sensitivity as measured through reaction norms.

**Selection for Robustness Traits** Livestock robustness can be defined as “the ability to combine a high production potential with resilience to stressors, allowing for unproblematic expression of a high production potential in a wide variety of environmental conditions” [123]. The classical problem with this ability is in genetic antagonisms between production traits and resilience [121] – natural selection is not powerful enough to maintain (or improve) animal robustness in intensive production systems; it must be supported by artificial selection. Genetic antagonisms can be neutralized by using adequate selection criteria to select for an adequate breeding goal. Earlier breeding goals were inadequate in this respect, as they did not include robustness traits [124]. Following Gjedrem [125], a breeding goal should include all heritable traits that have an impact on profitability – and mortality, morbidity, and lameness certainly do. Knol et al. [126] discuss this in detail and stress that “simplicity and straightforwardness of the breeding goal has to be weighted against completeness and complexity.” Such traits can be included in the profit equation for pig production [127]; this provides marginal economic values, required for inclusion in the breeding goal.

The various strategies for genetic improvement of piglet vitality and survival, leg weakness and longevity, stress sensitivity, and disease resistance are summarized in [128]. These are hard-to-measure traits, mostly categorical with low incidences and relatively low heritabilities, so that large data volumes from adverse environments are required for meaningful breeding value estimation [129]. Their genetic improvement has benefited considerably from BLUP and will benefit just as much from MAS [130, 131]. [Figures 9–11](#) show



**Pig Breeding for Increased Sustainability. Figure 9**

Simultaneous genetic trends of growth rate and leg soundness in eight pig lines. Each trait was scaled by the standard deviation of its estimated breeding values; all trend lines for each trait were forced through the same origin in 2000 to make the trends comparable across lines

realized genetic improvement of leg soundness and mortality traits, coinciding with improvement of production performance, in eight pig lines.

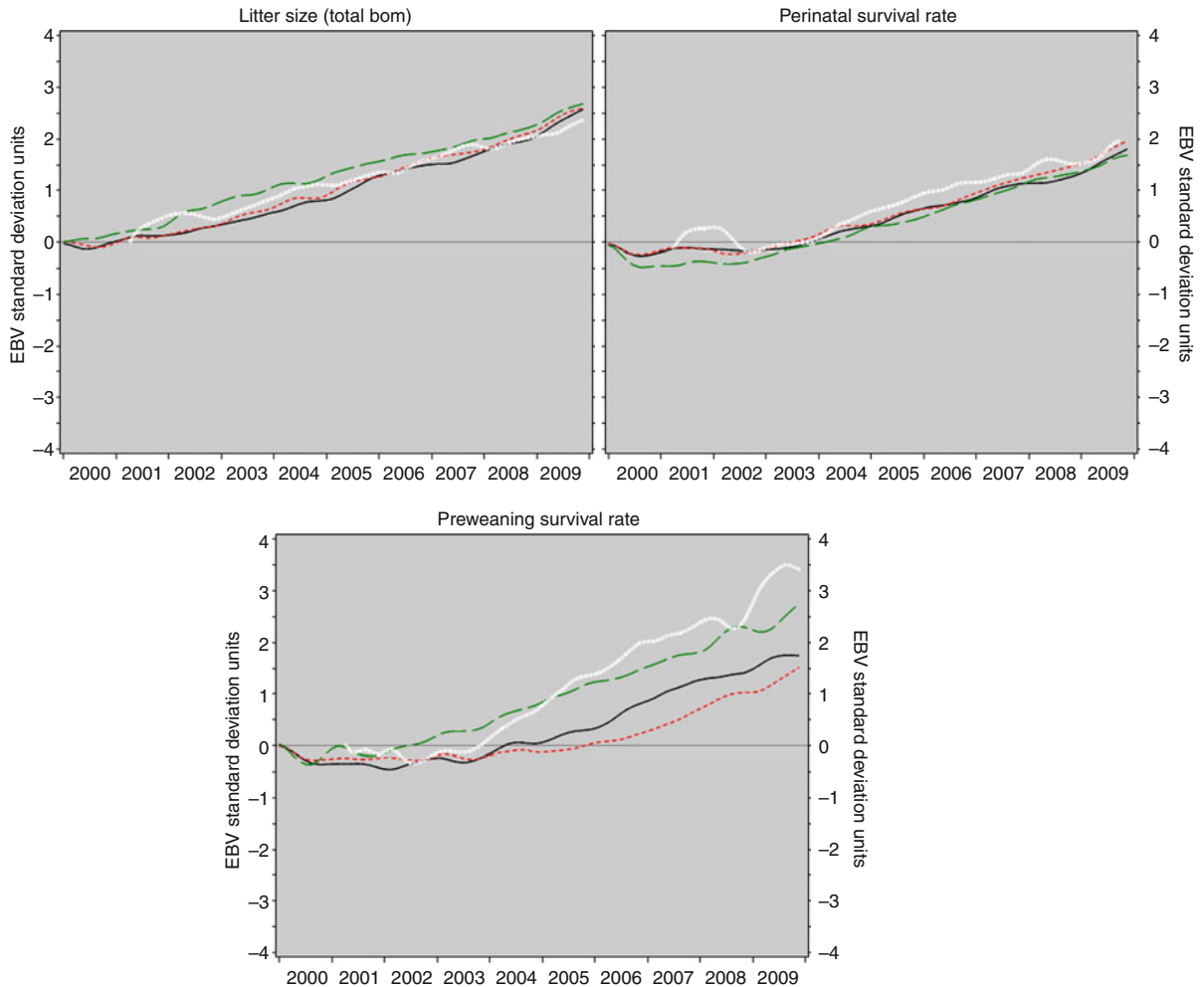
SVC [117] write: “The criterion for selecting animals for use in breeding has been an increase in economic performance and this has often not coincided with improved animal welfare. Hence the term *genetic improvement* is misleading since, in some cases, the changes are not improvements for the animal but may make the life of the animal more difficult.” The genetic trends of Figs. 9–11 show that this can be overcome: antagonistic correlations do exist, but genetic improvement can be achieved in production traits and robustness traits simultaneously – neutralizing genetic antagonisms by adequate selection. This effect will depend on the emphasis given to each trait in the breeding goal of each line.

**Reaction Norms** When progeny of specific sires are (1) identified as such, (2) spread across a wide environmental range (usually through artificial insemination), and (3) recorded for a production trait, their production performance can be regressed on a descriptor of the environment (e.g., a herd-year-season effect). This produces a positive slope overall:

better environments lead to better production. When the regression is performed separately for sire progeny groups, and if there is genetic variation in environmental sensitivity of the trait’s production potential, regression lines are produced (“reaction norms” in population genetics) with different intercepts and slopes for different sires. The intercepts are equivalent to the conventional estimated breeding values for the trait. The slopes quantify an animal’s requirements for environmental support of its genetic potential; they detect robustness as defined above.

Environmental sensitivity in pigs was analyzed this way in [132, 133]. Friggens and Van der Waaij [134] discuss how selection for increased production levels (i.e., for high reaction norm intercepts) will cause a gradual increase of environmental sensitivity (i.e., of the slopes). This was confirmed [133] in terms of a strongly positive genetic correlation between intercept and slope of their reaction norms for litter size. The slopes have a very low heritability in that data, so the increase of environmental sensitivity would be very slow. This presents another example of genetic antagonisms, which can be neutralized by including both the intercept and the slope of the reaction norm of each production trait in the breeding goal and in the





**Pig Breeding for Increased Sustainability. Figure 11**

Simultaneous genetic trends of litter size, perinatal survival rate, and preweaning survival rate in four pig damlines (same formatting as in Fig. 9)

challenges [...] an animal that maintains milk production by suppressing its immune function is clearly less robust than an animal that maintains milk production by reducing growth rate” [134]. Which means that (1) it becomes interesting how the environmental sensitivity rates are correlated among traits, and (2) breeding goals must be designed and monitored with care, as argued above.

### Deprivation

Intensive housing and management systems tend to compromise the expression of instinctive behavior

patterns (“motivations” [135]) of the pig: foraging, rooting, and exploring in all age classes [117], and nest-building in sows [136] – because intensive housing environments do not provide the required space or substrates. Something similar holds for deprivation of social contact in individually housed sows [137], where the “required substrate” is other pigs. For more detail see [138]. This results in frustration, leading to apparently irrationally redirected behavioral outlet functions: stereotypies or apathy occur when the animal is severely or chronically frustrated [139–141]. Mason and Bateson [142] describe this situation as “internal



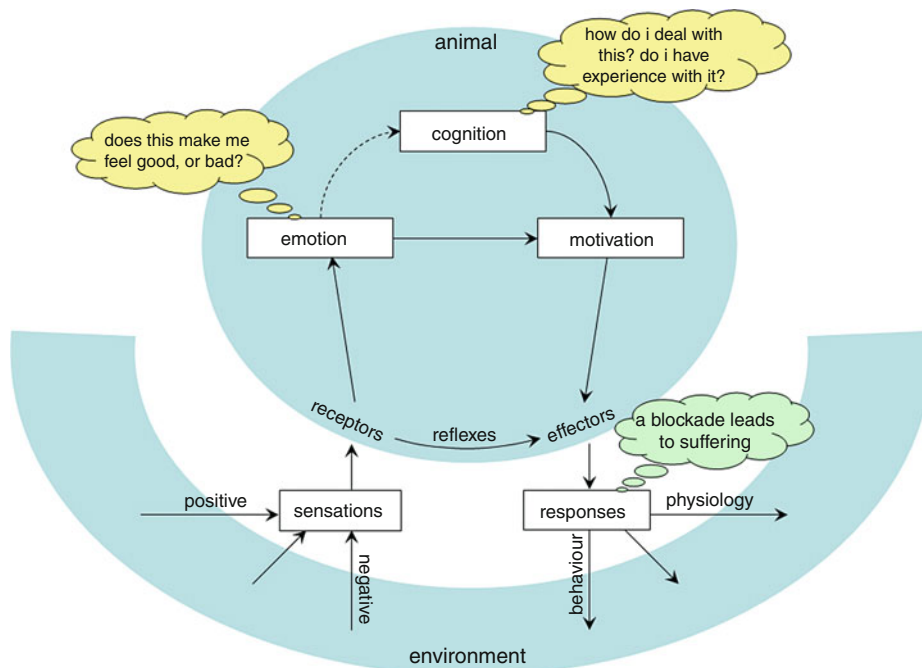
states of deprivation [...] leading to sustained high motivations that the [animal] cannot reduce: it cannot perform the activity that would result in negative feedback.” Another outlet function is wide-sense agonistic behavior such as tail biting [143, 144]; see the section on “[Harmful Social Behavior](#).”

In sentient species, such frustration leads to suffering. A sentient animal has been described [145] as an animal that has the capacity to suffer when it learns that it is unable to cope with stress: “it may fail to cope, either because the stress is too [intense], or because [the animal] is constrained in such a way that it is prevented from doing what it feels necessary to relieve the stress.” This leads to anxiety or depression. [Figure 12](#) illustrates the interdependence of the various components.

Held et al. [138] write: “pigs used today for pork production seem to have retained many of the faculties of their wild ancestors, and may therefore be behaviourally, cognitively and emotionally at odds with the husbanded environment [...]. Understanding

the cognitive abilities, behavioural priorities and emotions of commercial pigs therefore lies at the very heart of improving their welfare.” Jensen [146] describes those “faculties of the wild ancestors” as “subtle [behavioural] differences between domestic and wild animals [...] attributed to modified stimulus thresholds, causing some behaviour patterns to become more common and others to be rarer during domestication.” Such changes in pig behavior are described in detail in [147, 148].

One of the options for resolving such cognitive and emotional conflicts in an intensive environment would be to further reduce those faculties of the wild boar, further modifying those stimulus thresholds and simplifying what the animal feels necessary to relieve its stress. Kruska [149] sees changes in behavior and brain size as “adaptations to the special ecological niche of domestication.” “Modern housing systems [have] a short history compared to the history of the pig as a domestic animal, and it is likely that adaptation has not kept pace with the intensification of pig



**Pig Breeding for Increased Sustainability. Figure 12**

Relationships between emotional, cognitive, and motivational information processing in sentient animals (Modified from [145] and [139])

husbandry” [150]; also, of course, because most housing systems were not designed to match behavioral needs.

**Neuroendocrinology** The neuroendocrinological aspects of *coping with stress* (as above) are studied in relation to human conditions such as mental depression [151]. This is a novel area for animal breeding, with legible summaries of the state of the art from [139, 152–154]. Chronic stress affects the hypothalamus-pituitary-adrenal (HPA) axis; it can cause a persistent overproduction of corticosteroids, which can disturb neurotransmitter systems by causing a chronic downregulation and/or an imbalanced activation of the various types of corticosteroid receptors in the brain. Probably independent of the HPA axis (but interacting with its corticosteroids) is the mediation of stress-induced stereotypic behavior by the mesolimbic pathway. Figure 13 presents a simplified model of these systems.

The sympathetic nervous system also influences corticosteroid production, and its interaction with the HPA axis leads to *active* versus *passive* coping strategies. These terms represent the extremes of a continuous distribution of animal-intrinsic behavioral flexibility. This has been described [154–157] as follows: active copers rely on stable conditions, show poor adaptation to changing conditions, and attempt to deal with any challenge through routines and behavior patterns that were successful previously, trying to remove the stressor or move themselves away from it. Passive copers show higher cognitive performance, thrive better in changing conditions, and face challenges by modifying their behavior to deal with a new stressor in its own way, aiming at reduction of the emotional impact of the stress.

This approach has been criticized as an attempt to fit a multidimensional system (reactivity to external factors) into a single (coping) dimension. Ramos and Mormède [158] mention three such dimensions (activity, emotionality, and aggressiveness), and present studies in rodents that quantify these through Principal Components Analysis and similar techniques. Obviously, control of the system through animal breeding would benefit from a detailed focus, so that the various dimensions can be brought under control independently from each other.

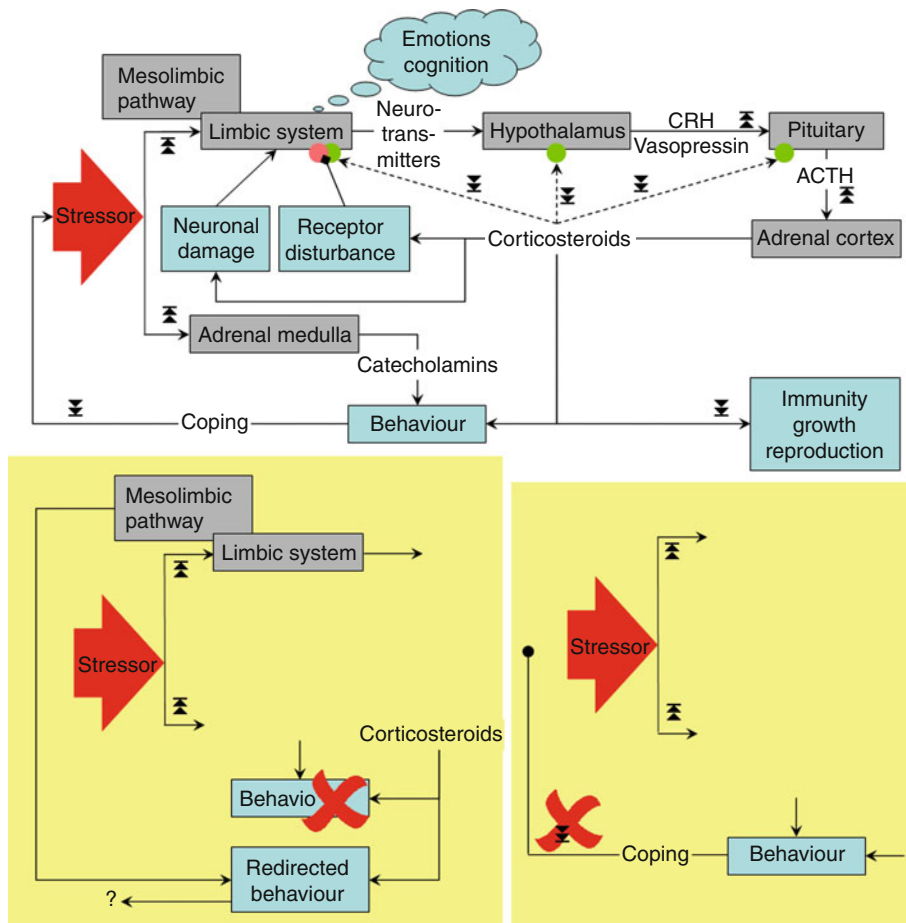
These traits are heritable (both coping strategies have natural selective advantages), and rodent populations have been successfully selected into either direction, using various selection criteria [159–161]. Veenema [154] found a higher stress susceptibility in passively coping than in actively coping mice. Her surmise is that this is due to differences in (1) perception of the stressor during acute stress, and (2) coping or habituation during chronic or repeated stress. Under changing conditions, passively coping mice may perform better in terms of dealing with the stressor. But exposure to chronic psychosocial stressors (i.e., living together with a dominant aggressive penmate) induced long-lasting increased activity of the HPA system in Veenema’s passively coping animals, which may cause mood disorders like anxiety and depression.

Karman [153] measured the effects of chronic stress (caused by 5 months of individual housing) in gilts that had previously been scored for coping strategy. She characterizes the two strategies in terms of differences in the regulation by the hypothalamus of the pituitary’s ACTH production: via CRH in passive copers, and via vasopressin in active copers. In both cases, ACTH levels and corticosteroid production are increased – which comes back to the disturbed neurotransmitter systems of above. Karman concludes that “individual housing is detrimental for the welfare of pigs, independent of coping strategy. Selection of coping strategy [...] will therefore not benefit the welfare of the animals.”

If Karman’s findings hold, the generic effects of chronic stress in pigs will not be resolved by breeding for an active coping strategy, which would seem to be the obvious approach with Veenema’s [154] mice.

This would shift the focus upstream in Fig. 13, aiming at a resolution, through animal breeding, of the effects of chronic stress on the limbic system areas (amygdala, hippocampus) that regulate the hypothalamus. De Kloet [156] makes a distinction between “(1) the core of the HPA axis with emphasis on dysregulations in the [hypothalamus] and (2) dysregulations in [...] stress inputs to the [hypothalamus] (e.g., from medial prefrontal cortex, hippocampus, amygdala, and brain stem) that are also targets for the stress hormones.”

Morris [162] notices that during chronic stress “the hippocampus may become impeded in its role in ‘shutting off’ HPA axis stress activity. This results in increased secretion of [corticosteroids] and, in



**Pig Breeding for Increased Sustainability. Figure 13**

A simplified model of the neuroendocrine aspects of unsuccessful coping with stress in mammals. *Top*: the full system with its negative-feedback loops that keep the levels of circulating corticosteroids under control. The flow starts with the occurrence of a stressor (red arrow). The  $\blacktriangleright$  symbol pointing up or down represents upregulation or downregulation, respectively, of the downstream activity. Green and pink circles represent different types of corticosteroid receptors; in the limbic system, these can be disturbed by persistently high corticosteroid levels, obstructing the negative feedback loop. The stressor is neutralized by coping behavior, which terminates its triggering of corticosteroid and catecholamine production. *Bottom*: two options for obstruction of this system. *Left*: an external factor prevents the expression of the required behavior, which leads to redirected behavior (e.g., stereotypies, agonistic behavior) as an outlet function with uncertain consequences. *Right*: an external factor keeps the stressor in place in spite of coping behavior, effectively blocking the stressor's downregulation that would normally follow from coping. In both cases, the stressor is not neutralized and the levels of circulating corticosteroids are out of control; this can result in a reduction of immunity, growth, and reproduction traits, and in damage to the limbic system which brings the system further out of control

a positive feedback cycle with negative consequences, ends up in damaging the hippocampus itself, thereby further reducing [its] ability to regulate the HPA axis. [...] This brain structure is both [i] centrally involved in the neural reaction to aspects of prolonged stress and

[ii] itself a target of chronic stress." Loijens [152] reports such findings in tethered sows.

The significant issue here is that the limbic system also regulates emotional and cognitive functions (Fig. 13) – which provides the conceptual connection

to Fig. 12. “Sustained activation of [corticosteroid] receptors in the hippocampus [...] may lead to an impairment of declarative (rule-) learning during high levels of chronic stress. [...] The amygdala [...] has been demonstrated to [influence] hippocampal plasticity and hence may be the central link between stress and declarative learning [...]. In farm animals this type of learning occurs during adaptation to [housing] facilities, milking regimes etc.” [139].

**Selection Objectives** “An array of stress-responsive genes has been identified,” including genes “related to structural differences in hippocampus of [passively and actively coping] mice. [...] These altered gene patterns can be postulated as markers for *predisposition* for stress-related disorders” [156]. Such gene patterns are being studied in pigs as well, exploring changes of gene expression in the hippocampus, amygdala, and/or frontal cortex due to early weaning and/or social isolation of piglets [163, 164].

“The evidence for a significant genetic contribution to stress responsiveness in vertebrates is overwhelming. [...] Given the complexity of the issues, there is no firm consensus as to whether modification of stress responsiveness can benefit an animal within an intensive rearing environment” [165].

In line with this, the main conclusion from the section on “**Deprivation**” would be that straightforward selection against specific behavior patterns such as stereotypies and apathy (which would not be difficult to record in intensively housed sows) may be counterproductive. In terms of Fig. 13, such selection might just remove the “redirected behavior” pathway that serves as an outlet for frustrated motivations, and as such forms the animal’s final way to deal with the load of the stressor. This would create a system under stress (with its negative consequences for homeostasis and production) without any security valve. D’Eath et al. [166] refer to such animals as *stoics* “because outward signs of suffering appear to be reduced” while the “root cause of the stereotypy” is not changed.

A second conclusion is that the active and passive coping strategies do not seem to offer a useful criterion for pig breeding.

It seems much more useful to look for the abovementioned “adaptations to the special ecological niche of domestication” [149] in terms of changes of the predisposition for stress-related disorders and of perception of the stressor, rather than a change of coping patterns. This would have to target the limbic system, which makes the task much more challenging.

From a very different point of view, Morris [162] quotes Sapolsky: “the body simply has not evolved the capacity or tendency to not secrete [corticosteroids] during a crisis,” and adds to this “in effect, evolution has only gotten so far.” Clearly, evolution could be moved on (into a more convenient direction) by a much focused targeting of the system that regulates the HPA axis, i.e., the limbic system again.

This amounts to a strategy to modify instinctive patterns so that the motivation for behavior that cannot be supported by the production system is reduced. This would “change the intensity of a behavioral response,” which is equivalent to domestication [119]; antipredator responses would be an obvious example. Adaptation of behavior through selection is then effectively an extension of 9,000 years of pig domestication: a process of reducing the animal’s drives for exploration, aggression, etc.

Such characteristics form the ultimate example of *hard-to-measure traits* in animal breeding. Genetic improvement will therefore logically focus on marker-assisted selection; phenotypic records will still be required in large volume for marker effect estimation, but recording can be scheduled on a project basis, and on other animals than selection candidates. As with every trait of livestock species studied up to now, very large numbers of genes are likely to be involved, which makes the candidate gene approach and its search for major genes not very promising – also because of the currently “limited basic knowledge about psychobiological dimensions underlying behavioral trait variability, and the availability of reliable and meaningful measures of these [...] free from environmental influences” [167]. The *quantomics* approach that was put forward at [www.quantomics.eu](http://www.quantomics.eu) seems to offer more power to bring such a system under control: it should “provide the tools to identify rapidly the causative DNA variation underpinning sustainability in livestock, and [...] exploit high-density genomic

information.” Essentially, making use of large numbers of anonymous markers (as in genomic selection) to identify functional elements and their connection to the phenotype (as in QTL studies).

**Dominance Aggression** Another issue of intensive production systems is avoidance of dominance aggression, particularly when pigs are being mixed into new groups [168, 169] – confinement does not allow for escape from aggressors. Such behavior has significant genetic components [170], possibly but not very clearly connected to coping strategies. The main factor that would complicate selection against such behavior is the difficulty of data recording, which makes marker-assisted selection an interesting option again. QTL associated with such behavior are in the process of being discovered [171, 172].

The ethical aspects of all this are considered in the section on “[Ethical Aspects](#).”

### Avoidance of Invasive Treatments

There is much societal drive to reduce painful treatments like castration and tail docking of piglets. The relevant issues are then the genetic options to reduce (1) boar taint and (2) tail biting – these form the reasons why those treatments are performed.

**Boar Taint** Piglet castration needs to be carried out under anesthesia in Norway and Switzerland. The German pig production sector has been recommending castration with analgesia since 2009, and aims at castration-free production on the longer term. The Dutch pig sector aims at castration-free production by 2015; leading Dutch retailers have decided to stop the sales of meat from castrated pigs starting 2011. A logical extrapolation is that by 2020 the European pig production sector will leave most of its male pigs uncastrated. While this provides progress in animal welfare (apart from aggression among entire male penmates), and an advance in gross profitability because entire males grow more efficiently than castrates [173, 174], it causes logistical and technical challenges because of boar taint.

Boar taint is an unpleasant odor of pig meat (occurring in roughly 3–10% of entire males) caused by several chemical components, most importantly

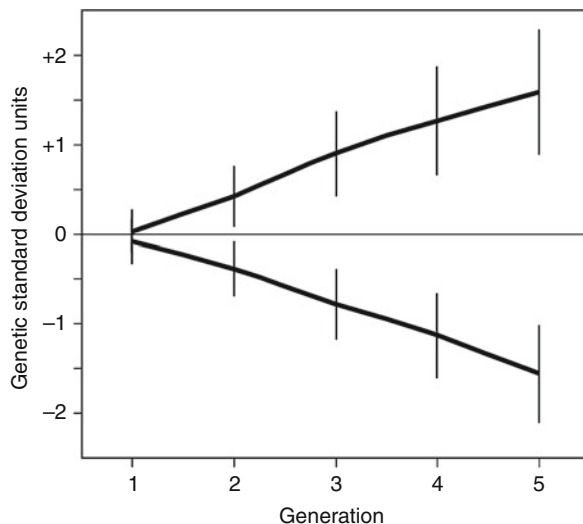
androstenone (a sex hormone) and skatole (a metabolite of the gut microflora). The tissue concentrations of both components are variable, line specific, and heritable [175–177], and the genes that influence them are gradually being identified [178, 179] so that it is feasible to select pigs for reduced boar taint levels. Such genetic improvement is a crucial element of the sustainability of sector-wide non-castration: although it is largely uncertain how consumers will react to increased amounts of meat from entire male pigs appearing on the market, boar taint incidence will have to be reduced to manageably low levels to avoid situations where consumer demand for pig meat collapses, or where the processing industry shifts to imported meat from castrates.

Because androstenone is a sex hormone, attempts to reduce its circulating levels may affect similar hormones such as testosterone and the estrogens, with a negative impact on male and female fertility, respectively. This must be counteracted through “balanced breeding” [177], i.e., by simultaneous selection for the relevant fertility traits.

**Harmful Social Behavior** SVC [117] write about “tail injury caused by biting. Although the motivation of the pig which bites the tail is likely to be investigation, manipulation and perhaps feeding rather than aggression, the consequence for the bitten pig is serious. Bitten tails may attract further biting so that the injury is to the abdomen at the base of the tail after the tail itself has been bitten off.” Tail biting is a form of “harmful social behaviour” [180] (HSB; “social” because it involves other pigs) in intensive animal production [181, 182], possibly a form of redirected outlet behavior (as in the section on “[Deprivation](#)”) in the absence of rooting substrate such as straw [183]. Other forms are vulva biting in group-housed sows [184], piglet savaging by sows [185], and feather pecking in poultry. Feather pecking and subsequent cannibalistic actions form a major problem in poultry production [186]. In practice, this vice is prevented by beak trimming of the potential actor, whereas tail biting in pigs is prevented by tail docking of the potential recipient; both at a very young age. Both treatments compromise animal welfare, but not performing them may do so too – a circular lose–lose situation to be broken.

Su et al. [187] selected laying hens for increased or reduced feather pecking incidence. Figure 14 shows the genetic trends in their first five generations. The selection trait was the number of pecking bouts delivered by a bird during 3 h, recorded by examining video footage of 250 birds each generation. Behavioral traits are notoriously time intensive to record and therefore a prime candidate for marker-assisted selection (phenotypic records will still be required in large volume for marker effect estimation, but recording can be scheduled on a project basis and on other animals than selection candidates). Motivated by this, “a major dominant allele affecting the [feather pecking] behavior” was identified in the eighth generation of the high-incidence line [188]. More genes with significant associations to feather pecking were found in similar selection lines [189]. Likewise, Quilter et al. [190] report on a search for QTL associated with piglet saving behavior in sows.

Social behavior traits involve a recipient and an actor; genetic evaluation should take both into



**Pig Breeding for Increased Sustainability. Figure 14** Genetic trends of feather pecking incidence in divergent laying-hen selection lines (Data from [187]). The data for each line have been scaled by the within-line genetic standard deviation. Vertical bars represent one standard error each side of the mean value

account. Ellen et al. [191] notice that “cannibalism [...] differs from conventional breeding traits because it depends on social interactions [...]. Selection strategies [...] should consider both the direct effect of an individual on its own survival and the social effect of the individual on the survival of its group members (the so-called associative effect).”

In group housing, an individual’s phenotype for any trait (growth rate, mortality, etc.) is influenced by its own direct breeding value for the trait and the associative breeding values of its penmates, positive or negative. Hence, if associative effects are significant for the trait, they should be part of the analysis in breeding value estimation – to effectively equip the social environment with a pedigree structure and capture it in more statistical detail. This principle has been worked out in detail for growth rate and feed intake in growing pigs [192, 193]; associative effects contributed the majority of heritable variance in these traits. Intuitively, the same would hold for mortality rates due to cannibalistic HSB.

Muir and Craig [194] discuss *group selection* in laying hens, where “hens of each sire family were housed as a group in a multiple-bird cage and selected as a group” for egg production and survival rate. These hens were housed intensively and were not subjected to beak trimming, allowing for unconstrained expression of HSB. After seven generations of selection, in group housing, the group selection line showed a 20% mortality rate at 58 weeks, compared to 54% in an unselected control line and 89% in a related commercial line (which was selected for egg production and survival rate in individual housing; Muir WM, personal communication, 2010). Plumage scores revealed significantly less HSB in the group selection line. Egg production in group housing was highest in the group selection line.

These authors conclude that this approach “is effective in improving [welfare] of layers in a relatively short period of time without sacrificing productivity. The way for commercial breeders to develop birds that do not need beak trimming is clear.” A study of the endocrine and immune system traits in these lines [195] concluded that “group selection altered the chickens’ physiological homeostasis which is reflected in the

line's unique coping ability with intensified domestic environments." But the physiological effects of this selection approach are cell-specific [196].

Gunsett [197] describes an application of group selection in pig breeding. He mentions advantages in terms of profitability (improved image of intensive production, increased stocking density, reduced abattoir penalties for damaged carcasses) and animal welfare: (1) reduced incidence of damaged carcasses, i.e., of pigs injured due to HSB, (2) reduced mortality rates, and (3) more docile behavior. He also mentions practical difficulties with the implementation of the method, due to reduced selection intensity and increased rate of inbreeding. Hence the program was changed to a system where the direct and associative breeding values were estimated in an extended BLUP approach [198], which was formally worked out by Bijma [199]. Group selection is then not required anymore. These principles were applied to HSB and its resulting mortality rates in laying hens, leading to the conclusion that "including associative effects in the model will give substantially higher heritable variation than when using the conventional direct effects model [...]; prospects for reduction of mortality using the direct-associative effects model are good [...]; selection targeting both direct and associative effects is expected to substantially reduce one of the major welfare problems in egg production" [191].

The extension to tail biting in pigs is obvious: Breuer et al. [180] estimated heritabilities of tail biting in two populations (both with an actor incidence of about 3%) at 0.00 and 0.05 on the observed scale, and bravely conclude that "it would be possible to develop a selection index to reduce [...] tail-biting behavior through selective breeding" – but any statistical method that delivers substantially higher heritable variation would be very useful here.

Muir and Craig [194] conclude from their selection results that "because group selection is shown to improve well-being in multiple-bird cages, alternatives such as redesigning cage environments, or housing such as floor pens or free ranges, may not be needed." Likewise, Conington et al. [200] write "breeding animals to adapt to their environment, rather than focus on changing environments to match new genotypes

(such as altering housing and cubicle design) can minimize the mismatch between them" [200]. Such statements are under debate, following the argument that it would be preferable to adapt the production system to the animal, rather than vice versa, which goes back to Faure [201]. This is in clear conflict with domestication in general, which has always attempted to adapt animals to captivity systems, sometimes by considerable force. The ethical aspects are dealt with in the next section.

### Ethical Aspects

Farm animals are there to produce food that people need and want. With an increasing human population and its worldwide purchasing power, there is an equally increasing demand (what people want) for animal products. One of the options of dealing with this increasing demand is to disapprove of it, arguing that the world's carrying capacity does not allow for it [202] and that everyone (particularly in the developed world) should eat less animal products. This raises one of the many arguments in an ethical discussion: to what extent it can be justified to deny people (particularly in the developing world) what they clearly want.

Another (nonexclusive) option is to look for technological solutions. Without doubt, intensive systems will be the norm in worldwide pig and poultry production of the 2020s: in USA, Brazil, the Middle East, Russia, and China – where the enforcement of extraneous norms and regulations is difficult and arguably ethically unjustified. Sections "[Robustness](#)" to "[Avoidance of Invasive Treatments](#)" show that pig breeding can produce genotypes that are better equipped to fare well in such systems. This would also enhance profitability in such systems (due to lower mortality and morbidity rates), leading to increased worldwide sustainability. The pragmatic approach would then be to accept this and aim at adapting the pig species to such conditions. There is much debate about this proposition.

Hörning [203] writes: "In general it seems ethically dubious when behavioral problems of intensive production must be reduced by breeding, rather than by

changing the management-related causes. On the other hand, selection for certain behavior patterns in extensive production conditions has been recommended by some livestock ethologists: [...] maternal behavior of sows in loose housing systems [...] or against feather pecking in laying hens in alternative housing systems” (translated).

In line with this, the genetic adaptation of goats, sheep, and beef cattle to marginal (mountainous, wetland, arid) extensive conditions is much explored [204, 205]. Those conditions often lead to severe and persistent violation of the freedom from thirst, hunger, thermal discomfort, parasites, and disease – the breeding of animals that withstand such harsh low-input conditions evokes Kojak’s maxim *to survive is a lousy way to live*. The justification of such adaptation strategies is commonly phrased in terms of the economical and/or cultural importance of livestock for such marginal areas (see the section on “**Biodiversity**,” and more specifically pp. 405–419 of [5]), which entirely overlooks animal welfare.

By contrast, animal welfare problems in intensive production systems center around robustness and behavioral deprivation. Strategies to reduce deprivation problems by animal breeding (as outlined in the section on “**Deprivation**”) meet with criticism of an ethical nature, as from Hörning [203], above.

The issue is if *genetic adaptation of livestock species* to the violation of their welfare by thirst, hunger, thermal discomfort, parasites, and disease in extensive conditions (for economic and cultural reasons) would be ethically justified, whereas genetic adaptation to behavioral deprivation in intensive systems (for economic reasons) would be wrong. If it is justified to select poultry against feather pecking in alternative housing systems, the question is valid why such selection in battery cages would be wrong.

There are two elements here: (1) the production system as such and (2) the process of adapting animal species to it through artificial selection. These are difficult to separate, because the argumentation is partly circular.

“Since biology appears to impose few limitations on what is possible, changing the animal to suit the environment raises the question of the ethical acceptability of the environment” [166]. The underlying notion here is that the environment may be intrinsically wrong. Nevertheless, *successful* adaptation of an animal species to any production system would make that system acceptable from the point of view of animal welfare – for that particular species, in that particular system. For anyone who finds such systems unacceptable a priori, such adaptation is therefore undesirable: the circular argument appears here. A housing system that is deemed unacceptable a priori, without taking animal welfare into account, can only make sense from a human perspective.

**Artificial Selection** Domestication through artificial selection is a human activity, and therefore subject to ethics. By contrast, natural selection just happens.

Natural selection has adapted species to previously hostile conditions: freezing (Antarctic icefish, *Dissostichus mawsoni*), molten sulfur (western Pacific tonguefish, *Symphurus thermophilus*), cobra venom (mongoose, *Herpestes ichneumon*), high CO<sub>2</sub> and low oxygen levels (naked mole rat, *Heterocephalus glaber*), compression (sperm whale, *Physeter macrocephalus*), drought (Arabian camel, *Camelus dromedarius*) and crowding (Mexican free-tailed bat, *Tadarida brasiliensis*), among many others. Natural selection is also adapting *Sus scrofa domesticus* to intensive housing conditions – this is happening now, but so slowly that it is very difficult to notice. Artificial selection can do the same, much faster.

Like natural selection, domestication used to be a slow process – its resulting changes were hardly noticeable with a normal human time horizon. These changes have accelerated considerably since the 1980s, to the extent that they are now measurable within, say, a decade (as in Figs. 9–11) – and many people feel uncomfortable with this. Despite education and popularization of science, the notion of evolution (i.e., genetic change) as a process that is actually taking place today is not widely appreciated. Judeo-Christian



culture regards species as fixed entities, which makes people resist a noticeable change of a species' instinctive repertoire because it is experienced as “unnatural” (which is exactly what it is: domestication, like every other aspect of civilization, is a deliberate move away from nature). This resistance is toward genetic change of the pig *as we know it*. Nine thousand years of domestication have reduced the cephalization ratio (brain size as a proportion of body size) of the domestic pig by 30–40% as compared to the wild boar, the same as the dog-wolf comparison [149, 206]; significantly, the limbic system is most affected, see the section on “[Selection Objectives](#).” There is no reason to think that this process has stopped – nor will any dog owner argue that the current situation is wrong. What many people resist is to notice such a process of change actually taking place, if it would be accelerated by more effective artificial selection procedures as in Belyaev's famous fox selection lines [207]. The pig *as we know it* represents one particular stage of an evolutionary continuum, much of which lies in the future. Because this stage is familiar, now, it is experienced as the “natural” one – which it is not: by definition there is no such thing as a natural domestic animal. Accordingly, FAWC [120] stress the distinction between “natural” and “normal” behavior in farm animals.

A common argument is that such further adaptations to the niche of domestication would reduce the animal to a means to an agricultural end, to a commodity – which “embodies an excessively instrumental view on living creatures” [208]; they would violate the animal's *integrity*, “making it in some way less complete than it was previously” [181].

**Integrity** “Would it be right to produce [...] a pig unable to feel pain and unresponsive to other pigs? [...] such a pig would not be able to suffer, and its use might lead to significant productivity gains. Someone arguing that [this] would be wrong, would not be able to argue thus on the grounds of animal suffering” [120]. Rather, such argumentation is typically phrased in terms of *integrity*. It is useful to distinguish between

two integration levels here: the individual animal, and the species as such.

One view is that violation of an individual animal's integrity is wrong; this is about production systems that keep animals in persistent pain, frustration, or fear – and [209] about breeding that predisposes animals to such conditions.

Quite another view is that it is wrong to breed animals that experience less pain, frustration, or fear in such production systems [210]. This is about species integrity: although the individual pig's integrity is less violated, it would be *less of a pig*, which sounds uncomfortable. Conversely, it can be argued that such a pig would be *less of a wild boar* (*Sus scrofa scrofa*) and therefore *more of a pig* (*Sus scrofa domesticus*) on the evolutionary continuum mentioned above. Importantly, the argument is not about animal welfare but about human values.

Appleby and Sandøe [211] analyze the various schools of philosophical thought on this issue “so that scientists may be more aware of the strengths and weaknesses of their own ideas about animal welfare.” Thompson [212] gives an overview of the same; one approach holds that it is important for animals to express their instinctive behavior motivations (see the section on “[Deprivation](#)”), as far as “they actually have these [...], but whether or not a given animal does or does not have these drives is immaterial. Or put differently, one cannot harm an animal by frustrating a [motivation] that it does not have. Because this view revolves around the [motivations] that individual animals actually have, it does not see anything problematic about producing animals that have different motivations.” This goes back to Rollin [213].

An opposing view is that such animals (1) “can be said to have been harmed, even if there is no corresponding adverse affect in terms of animal bodies or animal minds”; this (2) “would regard the use of genetic strategies to address welfare problems as morally problematic” [212]. Gavrell Ortiz [214] disagrees with point (1) but defends point (2) on the grounds of violated animal *dignity*, “even if the modification would improve the animal's welfare.”

All this reduces animal welfare to “a subset of human welfare, the animals’ preferences and [welfare] having relevance only to the extent that they are important to us” [215]. This was extended by Würbel [216]: “it is [. . .] important to distinguish between our intention to protect animals (which may be partly selfish) and true animal protection. Animal protection is *ethically* justified by our own human values. What animals need for their protection, however, needs to be justified *biologically* by values that apply to the animals. Only by acknowledging this distinction will we arrive at an ethical and legal framework that satisfies our ethical claims as well as doing justice to the animals” (our emphasis).

The conclusion from this section on “[Animal Welfare](#)” is simple and uncomfortable. Intensive pig production systems will expand considerably, particularly in the developing world. Adapting the species to such conditions is technically challenging but feasible; it will improve animal welfare. Argumentation against this serves human moral values, and not animal welfare.

### Future Directions

To repeat from the end of the [Introduction](#): *Sustainability will always be a matter of more or less: it can never be an absolute goal.* This can now be made more concrete in terms of conflicting concerns about the various targets of sustainable production (people, pigs, planet, profit), as follows.

De Boer and Cornelissen [217] evaluated three laying-hen housing systems for the sustainability indicators economic performance, ammonia emission, energy use, animal welfare, farmer welfare, and egg quality. These authors notice conflicts such as “improvement of farmer welfare is difficult to achieve in animal-friendly [systems], because unfavorable thoracic dust concentrations [. . .] are a direct result of the presence of litter.” With equal weighting to each indicator, the battery cage ranked considerably better for overall sustainability than deep-litter and aviary systems.

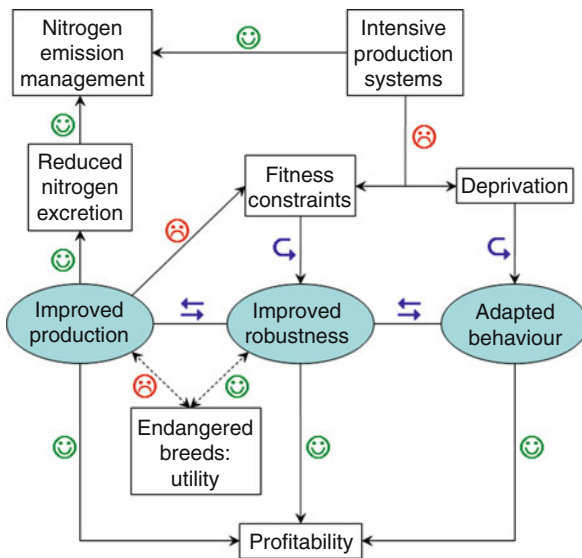
Likewise, three scenarios for enhanced sustainability of pig production were studied [218, 219] focusing on (1) animal welfare, (2) pollution, or (3) product quality and safety. Among the reported conflicts are

a higher contribution to acidification, a higher greenhouse gas emission, and higher production costs in (1) the animal welfare scenario than in (2) the pollution scenario. “Ranking between the different aspects of sustainability may [. . .] differ between different people and over time. How to evaluate [them] is mainly a political question, and legislation and political decisions can easily change the ranking of the scenarios.” This is about production systems, differing in animal management, housing, and feed production strategies. The breeding goals specified in that study were not dramatically different among the scenarios: (1) lean tissue growth rate (LTGR), feed intake, mothering ability, and longevity; (2) LTGR and feed efficiency; and (3) LTGR and meat quality – sets that occur alongside each other in any transnational breeding program, see [Figs. 9–11](#) in the section on “[Selection for Robustness Traits](#),” and also [Fig. 5](#) in [107].

However, the inclusion in breeding programs of specific sustainability targets (biodiversity, pollution, animal welfare) will create conflicts that need careful prioritization, as illustrated in [Fig. 15](#). Priorities can be based on economic approaches such as shadow prices (as illustrated for nitrogen excretion in the [Pollution](#) section), or benefit functions for animal welfare traits [220]. But the outcome must be a political compromise, and as such will change over time.

A recurring theme of this chapter is: *this technology is statistically demanding.* It follows that sustainable animal breeding equals high-tech animal breeding – just as sustainable animal production must involve precision farming [221] to overcome its inherent conflicts. Earlier breeding programs delivered at lower sustainability levels not only because of incomplete breeding goals (focusing on narrow-sense profitability) but also because the required data recording and processing methodology was not available.

Future directions will have to be set by the production sector and the society that it is part of. The commercial pig breeding sector does not have its own agenda: breeding goals are set based on what the market for breeding stock demands, and those demands are influenced by society – e.g., through legislation, or market regulation, around pollution and animal welfare. This chapter shows that the genetic technology to meet such demands is available, and can be exploited.



**Pig Breeding for Increased Sustainability. Figure 15**

Elements of pig breeding goals (blue ovals) and their relationships with sustainability issues. The symbols ☹️ and 😊 indicate (un-)favorable influences on the downstream element; ↻ indicates that solution of the upstream issue requires the downstream element; ↔ indicates that a balance must be established. Dotted arrows characterize the sustainability issue in terms of breeding goal elements

## Acknowledgments

Thanks are due to Gé Backus, Andrea Doeschl-Wilson, Eildert Groeneveld, Irene Hoffmann, Robert Hoste, Bas Kemp, Cees de Lange, Herveleine Lenoir, Asko Mäki-Tanila, Elzbieta Martyniuk, Dave McLaren, Marnie Mellencamp, Patrick Morel, Bill Muir, Candido Pomar, Rainer Roehe, Lotta Rydhmer, Montse Torremorell, Simon Turner, Eldon Wilson, and Hanno Würbel.

## Bibliography

- Gamborg C, Sandøe P (2003) Breeding and biotechnology in farm animals: ethical issues. In: Levinson R, Reiss MJ (eds) Key issues in bioethics: a guide for teachers. RoutledgeFalmer, London, pp 133–141
- Gamborg C, Sandøe P (2005) Applying the notion of sustainability – dilemmas and the need for dialogue. In: Gunning J, Holm S (eds) Ethics, law and society. Ashgate, Aldershot, pp 123–130
- Elkington J (1999) Cannibals with forks – the triple bottom line of 21st century business. Capstone, Oxford
- Knap PW, Neeteson-Van Nieuwenhoven AM (2006) Private and public roles in conservation. In: Options and strategies for the conservation of farm animal genetic resources: report of an international workshop and presented papers (CD-ROM). CGIAR Biodiversity International, Rome, pp 62–67
- FAO (2007) In: Rischkowsky B, Pilling D (eds) The state of the world's animal genetic resources for food and agriculture. Food and Agriculture Organization, Rome. ISBN 978-92-5-105762-9
- Rege JEO (1999b) The state of African cattle genetic resources. 1: Classification framework and identification of threatened and extinct breeds. Anim Genet Res Info 25:1–25
- Rege JEO, Gibson JP (2003) Animal genetic resources and economic development: issues in relation to economic valuation. Ecol Econ 45:319–330
- Tisdell C (2003) Socioeconomic causes of loss of animal genetic diversity: analysis and assessment. Ecol Econ 45: 365–376
- Gandini GC, Oldenbroek JK (2007) Strategies for moving from conservation to utilisation. In: Oldenbroek JK (ed) Utilisation and conservation of farm animal genetic resources. Wageningen Academic Publishers, Wageningen, pp 29–54
- Simianer H (2005) Decision making in livestock conservation. Ecol Econ 53:559–572
- Gandini GC, Villa E (2003) Analysis of the cultural value of local livestock breeds: a methodology. J Anim Breed Genet 120:1–11
- Ollivier L (2005) Economic relevance of animal diversity conservation. In: Rosati A, Tewolde A, Mosconi C (eds) Animal production and animal science worldwide. Wageningen Academic Publishers, Wageningen, pp 271–279
- Hill WG, Zhang XS (2009) Maintaining genetic variation in fitness. In: Van der Werf JHJ, Graser HU, Frankham R, Gondro C (eds) Adaptation and fitness in animal populations. Springer, Berlin, pp 59–82
- Nicholas FW (2009) Discussion. In: Van der Werf JHJ, Graser HU, Frankham R, Gondro C (eds) Adaptation and fitness in animal populations. Springer, Berlin, pp 233–234
- Dempfle L (1990) Conservation, creation, and utilization of genetic variation. J Dairy Sci 73:2593–2600
- Megens HJ, Crooijmans RPMA, SanCristobal M, Hui X, Li N, Groenen MAM (2008) Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: differences in microsatellite variation between two areas of domestication. Genet Sel Evol 40:103–128
- Eding H, Bennewitz J (2007) Measuring genetic diversity in farm animals. In: Oldenbroek JK (ed) Utilisation and conservation of farm animal genetic resources. Wageningen Academic Publishers, Wageningen, pp 103–130
- San Cristobal M, Chevalet C, Haley CS, Joosten R, Rattink AP, Harlizius B, Groenen MAM, Amigues Y, Boscher M-Y, Russell G, Law A, Davoli R, Russo V, Désautels C, Alderson L, Fimland E, Bagga M, Delgado JV, Vega-Pla JL, Martinez AM, Ramos M, Glodek P, Meyer JN, Gandini GC, Matassino D, Plastow GS, Siggins KW, Laval G, Archibald AL, Milan D, Hammond K, Cardellino R (2006a) Genetic diversity within and between

- European pig breeds using microsatellite markers. *Anim Genet* 37:189–198
19. SanCristobal M, Chevalet C, Peleman J, Heuven H, Brugmans B, Van Schriek M, Joosten B, Rattink AP, Harlizius B, Groenen MAM, Amigues Y, Boscher MY, Russell G, Law A, Davoli R, Russo V, Désautels C, Alderson L, Finland E, Bagga M, Delgado JV, Vega-Pla JL, Martinez AM, Ramos M, Glodek P, Meyer JN, Gandini G, Matassino D, Siggens K, Laval G, Archibald A, Milan D, Hammond K, Cardellino R, Haley C, Plastow G (2006b) Genetic diversity in European pigs utilizing amplified fragment length polymorphism markers. *Anim Genet* 37:232–238
  20. Van Zeveren A, Peelman L, Van de Weghe A, Bouquet Y (1995) A genetic study of Belgian pig populations by means of seven microsatellites. *J Anim Breed Genet* 112:191–204
  21. Sollero BP, Paiva SR, Faria DA, Guimarães EF, Castro STR, Egito AA, Albuquerque MSM, Piovezan U, Bertani GR, Da San Mariante A (2009) Genetic diversity of Brazilian pig breeds evidenced by microsatellite markers. *Livest Sci* 123:8–15
  22. Souza CA, Ramayo Y, Megens HJ, Rodríguez MC, Loarca A, Caal E, Soto H, Melo M, Revidatti MA, De la Rosa SA, Shemereteva IN, Okumura N, Cho IC, Delgado JV, Paiva SR, Crooijmans RPMA, Schook LB, Groenen MAM, Ramos-Onsins SE, Pérez-Enciso M (2010) Porcine colonization of the Americas: a 60k SNP story. In: 9th WCGALP, Leipzig, Germany (Communication 0510)
  23. Alves E, Barragán C, Fernández AI, Rodríguez C, Silió L (2006) Success rate of genetic clustering of domestic and wild pigs as a function of the number of markers. In: 8th WCGALP, Belo Horizonte, Brazil (Communication 33–24)
  24. Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783–791
  25. Flury C, Weigend S, Ding X, Täubert H, Simianer H (2007) Haplotype kinship for three populations of the Goettingen minipig. *Genet Sel Evol* 39:159–179
  26. Petit RJ, El Mousadik A, Pons O (1998) Identifying populations for conservation on the basis of genetic markers. *Conserv Biol* 12:844–855
  27. Toro MA, Mäki-Tanila A (2007) Genomics reveals domestication history and facilitates breed development. In: Oldenbroek JK (ed) *Utilisation and conservation of farm animal genetic resources*. Wageningen Academic Publishers, Wageningen, pp 75–102
  28. Eding H, Crooijmans RPMA, Groenen MAM, Meuwissen THE (2002) Assessing the contribution of breeds to genetic diversity in conservation schemes. *Genet Sel Evol* 34:613–633
  29. Fabuel E, Barragán C, Silió L, Rodríguez MC, Toro MA (2004) Analysis of genetic diversity and conservation priorities in Iberian pigs based on microsatellite markers. *Heredity* 93:104–113
  30. Bennewitz J, Meuwissen THE (2005) A novel method for the estimation of the relative importance of breeds in order to conserve the total genetic variance. *Genet Sel Evol* 37:315–337
  31. Ollivier L, Foulley JL (2005) Aggregate diversity: new approach combining within- and between-breed genetic diversity. *Livest Prod Sci* 95:247–254
  32. Toro MA, Fernández J, Caballero A (2009) Molecular characterization of breeds and its use in conservation. *Livest Sci* 120:174–195
  33. Ciobanu DC, Day AE, Nagy A, Wales R, Rothschild MF, Plastow GS (2001) Genetic variation in two conserved local Romanian pig breeds using type 1 DNA markers. *Genet Sel Evol* 33:417–432
  34. Iannucelli N, Riquet J, Mercat MJ, Legros H, SanCristobal M, Lacoste A, Bidanel JP, Milan D (2006) Diversité génétique des populations porcines françaises dans les régions chromosomiques soumises à la sélection: projet DIVQTL. *Les Actes du BRG* 6:111–128
  35. Bonin A, Nicole F, Pompanon F, Miaud C, Taberlet P (2007) Population adaptive index: a new method to help measure intraspecific genetic diversity and prioritize populations for conservation. *Conserv Biol* 21:697–708
  36. Falconer DS, Mackay TFC (1996) *Introduction to quantitative genetics*, 4th edn. Longman, Harlow
  37. Draganescu C, Ghita E, Nagy A (2008) Note on the genetic history of the Romanian Saddleback (Bazna) pig breed conservation nucleus. *Archiva Zootechnica* 11:65–69
  38. Nomura T, Ieiri S, Yamashita J (2008) Effective size and inbreeding in closed breeding herds of pig. 1: Pedigree analysis. *Jpn J Swine Sci* 45:149–155
  39. Dikić M, Salajpal K, Karolyi D (2009) Biological characteristics of Turopolje pig breed as factors in renewing and preservation of population. DAGENE annual meeting, Zagreb, Croatia, [http://www.dagene.eu/annual\\_zagreb\\_2009/Dikic\\_et\\_al\\_DAGENE\\_2009.pdf](http://www.dagene.eu/annual_zagreb_2009/Dikic_et_al_DAGENE_2009.pdf)
  40. Welsh CS, Blackburn HD, Schwab C (2009) Population status of major US swine breeds. *J Anim Sci* 87(Suppl 3):160
  41. Tholen E, Staack J, Müller P, Ingwersen J (2010) Degree of endangerment of different German pig herdbook populations. In: 9th WCGALP, Leipzig, Germany (Communication 0264)
  42. Uimari P, Sevón-Aimonen ML, Strandén I (2010) Pedigree analysis of Finnish Landrace and Yorkshire pig populations. In: 9th WCGALP, Leipzig, Germany (Communication 0090)
  43. Villanueva B, Sawalha RM, Roughsedge T, Rius-Vilarrasa E, Woolliams JA (2010) Development of a genetic indicator of biodiversity for farm animals. *Livest Sci* 129:200–207
  44. Wang J (2005) Estimation of effective population sizes from data on genetic markers. *Philos Trans R Soc Lond B Biol Sci* 360:1395–1409
  45. Álvarez I, Royo LJ, Gutiérrez JP, Fernández I, Arranz JJ, Goyache F (2008) Relationship between genealogical and microsatellite information characterizing losses of genetic

- variability: empirical evidence from the rare Xalda sheep breed. *Livest Sci* 115:80–88
46. Tixier-Boichard M, Ayalew W, Jianlin H (2008) Inventory, characterization and monitoring. *Anim Genet Res Info* 42:29–44, Food and Agriculture Organization, Roma, Italy
  47. Toro MA, Rodriganez J, Silio L, Rodriguez C (2000) Genealogical analysis of a closed herd of black hairless Iberian pigs. *Conserv Biol* 14:1843–1851
  48. Aspi J, Roininen E, Ruokonen M, Kojola I, Vilà C (2006) Genetic diversity, population structure, effective population size and demographic history of the Finnish wolf population. *Mol Ecol* 15:1561–1576
  49. Meuwissen THE (2007) Operation of conservation schemes. In: Oldenbroek JK (ed) *Utilisation and conservation of farm animal genetic resources*. Wageningen Academic Publishers, Wageningen, pp 167–194
  50. Ollivier L, Alderson L, Gandini GC, Foulley JL, Haley CS, Joosten R, Rattink AP, Harlizius B, Groenen MAM, Amigues Y, Boscher MY, Russell G, Law A, Davoli R, Russo V, Matassino D, Désautés C, Fimland E, Bagga M, Delgado JV, Vega-Pla JL, Martinez AM, Ramos AM, Glodek P, Meyer JN, Plastow GS, Siggins KW, Archibald AL, Milan D, SanCristobal M, Laval G, Hammond K, Cardellino R, Chevalet C (2005) An assessment of European pig diversity using molecular markers: partitioning of diversity among breeds. *Conserv Genet* 6:729–741
  51. Simon DL, Buchenauer D (1993) Genetic diversity of European livestock breeds. EAAP publication 66, Wageningen Pers, Wageningen, Netherlands
  52. Gandini GC, Ollivier L, Danell B, Distl O, Georgoudis A, Groeneveld E, Martyniuk E, Van Arendonk JAM, Woolliams JA (2004) Criteria to assess the degree of endangerment of livestock breeds in Europe. *Livest Prod Sci* 91:173–182
  53. Lynch M, Conery J, Burger R (1995) Mutation accumulation and the extinction of small populations. *Am Nat* 146:489–518
  54. Theodorou K, Couvet D (2006) On the expected relationship between inbreeding, fitness, and extinction. *Genet Sel Evol* 38:371–387
  55. Bennewitz J, Meuwissen THE (2005) Estimation of extinction probabilities of five German cattle breeds by population viability analysis. *J Dairy Sci* 88:2949–2961
  56. Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM (2007) Recent human effective population size estimated from linkage disequilibrium. *Genome Res* 17(4):520–526
  57. Hayes BJ, Visscher PM, McPartlan HC, Goddard ME (2003) Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res* 13:635–643
  58. Amaral AJ, Megens HJ, Crooijmans RPMA, Heuven HCM, Groenen MAM (2008) Linkage disequilibrium decay and haplotype block structure in the pig. *Genetics* 179:569–579
  59. Du FX, Clutter AC, Lohuis MM (2007) Characterizing linkage disequilibrium in pig populations. *Int J Biol Sci* 3:166–178
  60. Harmegnies N, Farnir F, Davin F, Buys N, Georges M, Coppieters W (2006) Measuring the extent of linkage disequilibrium in commercial pig populations. *Anim Genet* 37:225–231
  61. England PR, Cornuet JM, Berthier P, Tallmon DA, Luikart G (2006) Estimating effective population size from linkage disequilibrium: severe bias in small samples. *Conserv Genet* 7:303–308
  62. Waples RS (2006) A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conserv Genet* 7:167–184
  63. Schwartz MK, Tallmon DA, Luikart G (1998) Review of DNA-based census and effective population size estimators. *Anim Conserv* 1:293–299
  64. Wang J (2009) A new method for estimating effective population sizes from a single sample of multilocus genotypes. *Mol Ecol* 18:2148–2164
  65. Meuwissen THE, Goddard ME (2007) Multipoint identity-by-descent prediction using dense markers to map quantitative trait loci and estimate effective population size. *Genetics* 176:2551–2560
  66. Vitalis R, Couvet D (2001) Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics* 157:911–925
  67. Abdallah JM, SanCristobal M, Chevalet C (2006) Estimation of effective size and migration rates in twelve European pig populations using non equilibrium model. In: 8th WCGALP, Belo Horizonte, Brazil (Communication 33–19)
  68. Baas TJ, Goodwin RN, Christian LL, Johnson RK, Robison OW, Mabry JW, Clark K, Tokach M, Henry S, Berger PJ (2003) Design and standards for genetic evaluation of swine seedstock populations. *J Anim Sci* 81:2409–2418
  69. Moeller SJ, Goodwin RN, Johnson RK, Mabry JW, Baas TJ, Robison OW (2004) The National Pork Producers Council maternal line national genetic evaluation program: a comparison of six maternal genetic lines for female productivity measures over four parities. *J Anim Sci* 82:41–53
  70. Gibson JP, Ayalew W, Hanotte O (2007) Measures of diversity as inputs for decisions in conservation of livestock genetic resources. In: Jarvis DI, Padoch C, Cooper HD (eds) *Managing biodiversity in agricultural ecosystems*. Columbia University Press, New York, pp 117–140. ISBN 978-0231136488
  71. Nichol ST, Arikawa J, Kawaoka Y (2000) Emerging viral diseases. *Proc Natl Acad Sci USA* 97:12411–12412
  72. Pybus OG, Rambaut A (2009) Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet* 10:540–550
  73. Covey S, Merrill AR, Merrill RR (1994) *First things first*. Simon and Schuster, New York. ISBN 0-684-80203-1
  74. Mendelsohn R (2003) The challenge of conserving indigenous domesticated animals. *Ecol Econ* 45:501–510
  75. Ollivier L, Foulley JL (2009) Managing genetic diversity, fitness and adaptation. In: Van der Werf JHJ, Graser HU, Frankham R,

- Gondro C (eds) *Adaptation and fitness in animal populations*. Springer, Berlin, pp 201–228
76. Simianer H, Reist-Marti SB, Gibson J, Hanotte O, Rege JEO (2003) An approach to the optimal allocation of conservation funds to minimize loss of genetic diversity between livestock breeds. *Ecol Econ* 45:377–392
  77. Bennewitz J, Eding H, Ruane J, Simianer H (2007) Selection of breeds for conservation. In: Oldenbroek JK (ed) *Utilisation and conservation of farm animal genetic resources*. Wageningen Academic Publishers, Wageningen, pp 131–146
  78. Barker JSF (2001) Conservation and management of genetic diversity: a domestic animal perspective. *Can J For Res* 31: 588–595
  79. Rege JEO (1999a) Characterisation and conservation of animal genetic resources: What is it about? In: Rege JEO (ed) *Economic valuation of animal genetic resources*. In: *Proceedings of an FAO/ILRI workshop*. International Livestock Research Institute, Nairobi, Kenya pp 22–24
  80. Delanotte M (2007) Un avenir en rose pour une race qui a parfois broyé du noir. L'Union Agricole 21 December 2007. [limousin.synagri.com/ca1/PJ.nsf/TECHJPARCLEF/08102](http://limousin.synagri.com/ca1/PJ.nsf/TECHJPARCLEF/08102)
  81. Woolliams JA, Berg P, Mäki-Tanila A, Meuwissen THE, Finland E (2005) Sustainable management of animal genetic resources. *Nordic Gene Bank Farm Animals*, Ås, Norway
  82. Rathje TA (2000) Strategies to manage inbreeding accumulation in swine breeding company nucleus herds: some case studies. *J Anim Sci* 79(Suppl 1):1–8
  83. Ollivier L, James JW (2004) Predicting the annual effective size of livestock populations. *Genet Res* 84:41–46
  84. Woolliams JA (2007) Genetic contributions and inbreeding. In: Oldenbroek JK (ed) *Utilisation and conservation of farm animal genetic resources*. Wageningen Academic Publishers, Wageningen, pp 147–165
  85. Kinghorn BP, Banks R, Gondro C, Kremer VD, Meszaros SA, Newman S, Shepherd RK, Vagg RD, Van der Werf JHJ (2009) Strategies to exploit genetic variation while maintaining diversity. In: Van der Werf JHJ, Graser HU, Frankham R, Gondro C (eds) *Adaptation and fitness in animal populations*. Springer, Berlin, pp 191–200
  86. Kremer VD, Newman SN, Kinghorn BP, Knap PW, Wilson ER (2009) Strategic management of pig genetics. In: *EAAP, Barcelona, Spain (Communication 13–25)*
  87. Nimbkar C, Gibson J, Okeyo M, Boettcher P, Soelkner J (2007) Sustainable use and genetic improvement. In: *Report of the scientific forum on animal genetic resources held during the international technical conference on animal genetic resources for food and agriculture*, Interlaken, Switzerland. Food and Agriculture Organization, Rome, Italy, pp 49–64
  88. Whittemore CT (2006) Development and improvement of pigs by genetic selection. In: Kyriazakis I, Whittemore CT (eds) *Whittemore's science and practice of pig production*, 3rd edn. Blackwell, Oxford, pp 184–262
  89. Knap PW (1990) The herdbook society/breed association. In: 4th WCGALP, Edinburgh, UK, 15:439–442
  90. Steinfeld H, Gerber P, Wassenaar T, Castel V, Rosales M, De Haan C (2006) *Livestock's long shadow: environmental issues and options*. Food and Agriculture Organization, Rome
  91. Brandjes PJ, De Wit J, Van der Meer HG, Van Keulen H (1995) *Environmental impact of animal manure management*. International Agricultural Centre, Wageningen
  92. Dourmad JY, Sève B, Latimier P, Boisen S, Fernández J, Van der Peet-Schwering C, Jongbloed AW (1999a) Nitrogen consumption, utilisation and losses in pig production in France, the Netherlands and Denmark. *Livest Prod Sci* 58:261–264
  93. Dourmad JY, Guingand N, Latimier P, Sève B (1999b) Nitrogen and phosphorus consumption, utilisation and losses in pig production: France. *Livest Prod Sci* 58:199–211
  94. Bourdon D, Dourmad JY, Henry Y (1995) Réduction des rejets azotés chez les porcs en croissance par la mise en oeuvre de l'alimentation multiphase, associée à l'abaissement du taux azoté. *Journées de la Recherche Porcine en France* 27:269–278
  95. De Lange CF, Gillis AM, Simpson GJ (2001) Influence of threonine intake on whole-body protein deposition and threonine utilization in growing pigs fed purified diets. *J Anim Sci* 79:3087–3095
  96. Buraczewska L, Świąch E, Le Bellego L (2006) Nitrogen retention and growth performance of 25 to 50 kg pigs fed diets of two protein levels and different ratios of digestible threonine to lysine. *J Anim Feed Sci* 15:25–36
  97. Pomar C, Hauschild L, Zhang GH, Pomar J, Lovatto PA (2010) Precision feeding can significantly reduce feeding cost and nutrient excretion in growing animals. In: Sauvart D, Van Milgen J, Faverdin P, Friggens N (eds) *Modelling nutrient digestion and utilisation in farm animals*. Wageningen Academic Publishers, pp 327–334
  98. Crocker AW, Robison OW (2002) Genetic and nutritional effects on swine excreta. *J Anim Sci* 80:2809–2816
  99. De Verdal H, Narcy A, Le Bihan-Duval E, Chapuis H, Bastianelli D, Mème N, Mignon-Grasteau S (2010) Selection for excretion traits in chicken. In: 9th WCGALP, Leipzig, Germany (Communication 0123)
  100. Cassandro M, Cecchinato A, Battagin M, Penasa M (2010) Genetic parameters of methane production in Holstein Friesian cows. In: 9th WCGALP, Leipzig, Germany (Communication 0837)
  101. Van der Peet-Schwering CMC, Jongbloed AW, Aarnink AJA (1999) Nitrogen and phosphorus consumption, utilisation and losses in pig production: the Netherlands. *Livest Prod Sci* 58:213–224
  102. Jones H, Warkup C, Williams A, Audsley E (2008) The effect of genetic improvement on emissions from livestock systems. In: 59th EAAP, Vilnius, Lithuania (Contribution 05-6)

103. Knap PW (1999) Simulation of growth in pigs: evaluation of a model to relate thermoregulation to body protein and lipid content and deposition. *Anim Sci* 68:655–679
104. Knap PW, Rauw WM (2008) Selection for high productivity in pigs. In: Rauw WM (ed) *Resource allocation theory applied to farm animal production*. CAB, Wallingford, pp 288–301
105. NRC (1998) *Nutrient requirements of swine*, 10th revised edn. National Academy Press, Washington, DC
106. Morel PCH, Wood GR (2005) Optimisation of nutrient use to maximise profitability and minimise nitrogen excretion in pig meat production systems. *Acta Horticulturae (ISHS)* 674: 269–275
107. Knap PW (2009) Voluntary feed intake and pig breeding. In: Torrallardona D, Roura E (eds) *Voluntary feed intake in pigs*. Wageningen Academic Publishers, Wageningen, pp 11–33
108. Teizzi S (1999) External effects of agricultural production in Italy and environmental accounting. *Environ Resour Econ* 13:459–472
109. Hartmann M, Hediger W, Peter S (2008) Reducing nitrogen losses from agricultural systems – an integrated economic assessment. *Schriften der Gesellschaft für Wirtschafts- und Sozialwissenschaften des Landbaus (Landwirtschaftsverlag, Münster, Germany)* 43:335–344
110. Fontein PF, Thijssen GJ, Magnus JR, Dijk J (1994) On levies to reduce the nitrogen surplus: the case of Dutch pig farms. *Environ Resour Econ* 4:455–478
111. Paul CJM, Ball VE, Felthoven RG, Grube A, Nehring RF (2002) Effective costs and chemical use in United States agricultural production: using the environment as a “free” input. *Am J Agr Econ* 84:902–915
112. Key N, McBride W, Mosheim R (2008) Decomposition of total factor productivity change in the US hog industry. *J Agr Appl Econ* 40:137–149
113. Fontein PF, Thijssen GJ, Magnus JR, Dijk J (1999) Optimal taxation for the reduction of nitrogen surplus Dutch dairy farms, 1975–1989. In: Mahendrarajah S, Jakeman AJ, McAleer M (eds) *Modelling change in integrated economic and environmental systems*. Wiley, Chichester, pp 273–296
114. Price R, Thornton S, Nelson S (2007) The social cost of carbon and the shadow price of carbon: what they are, and how to use them in economic appraisal in the UK. Department for Environment, Food and Rural Affairs, London
115. Wall E, Simm G, Moran D (2010) Developing breeding schemes to assist mitigation of greenhouse gas emissions. *Animal* 4:366–376
116. Hoste R, Puister L (2009) Productiekosten van varkens: een internationale vergelijking. LEI Wageningen UR, Den Haag, Netherlands. Report 2008-082
117. SVC (1997) The welfare of intensively kept pigs. Scientific Veterinary Committee of the European Commission, Brussels, Belgium. Report XXIV/B3/ScVC/0005/1997
118. Fraser D (2005) Animal welfare and the intensification of animal production. An alternative interpretation. Food and Agriculture Organization, Rome
119. Grandin T, Deesing MJ (1998) Genetics and behavior during handling, restraint and herding. In: Grandin T (ed) *Genetics and the behavior of domestic animals*. Academic, San Diego, CA, pp 113–144
120. FAWC (2009) *Farm animal welfare in Great Britain: past, present, future*. Farm Animal Welfare Council, London
121. Rauw WM, Kanis E, Noordhuizen-Stassen EN, Grommers FJ (1998) Undesirable side effects of selection for high production efficiency in farm animals: a review. *Livest Prod Sci* 56:15–33
122. Knap PW, Rauw WM (2009) Selection for high production in pigs. In: Rauw WM (ed) *Resource allocation theory applied to farm animal production*. CAB International, Wallingford, pp 210–229
123. Knap PW (2005) Breeding robust pigs. *Aust J Exp Agric* 45:763–774
124. Goddard M (2009) Fitness traits in animal breeding programs. In: Van der Werf JHJ, Graser HU, Frankham R, Gondro C (eds) *Adaptation and fitness in animal populations*. Springer, Berlin, pp 41–52
125. Gjedrem T (1972) A study on the definition of the aggregate genotype in a selection index. *Acta Agric Scand* 22:11–16
126. Knol EF, Duijvesteijn N, Leenhouwers JI, Merks JWM (2010) New phenotypes for new breeding goals in pigs. In: EAAP Heraklion, Greece (Communication 02-2)
127. Knap PW (2009a) Robustness. In: Rauw WM (ed) *Resource allocation theory applied to farm animal production*. CAB International, Wallingford, pp 288–301
128. Rydhmer L, Lundeheim N (2008) Breeding pigs for improved welfare. In: Faucitano L, Schaefer AL (eds) *Welfare of pigs from birth to slaughter*. Wageningen Academic Publishers, Wageningen, pp 243–270
129. Newman S, Wang L, Anderson J, Casey D (2010) Utilizing crossbred records to increase accuracy of breeding values in pigs. In: 9th WCGALP, Leipzig, Germany (Communication 0632)
130. Henryon M, Sørensen AC, Berg P, Nielsen B (2010) Breeding pigs for resistance to disease is difficult even with genomic selection. In: 9th WCGALP, Leipzig, Germany (Communication 0854)
131. Onteru SK, Fan B, Garrick DJ, Stalder KJ, Rothschild MF (2010) Whole-genome association analyses for sow lifetime production, reproduction and structural soundness traits using the PorcineSNP60 beadchip. In: 9th WCGALP, Leipzig, Germany (Communication 0273)
132. Hermes S, Huisman AE, Luxford BG, Graser HU (2006) Analysis of genotype by feeding level interaction in pigs applying reaction norm models. In: 8th WCGALP, Belo Horizonte, Brazil (Contribution 06-03)

133. Knap PW, Su G (2008) Genotype by environment interaction for litter size in pigs as quantified by reaction norms analysis. *Animal* 2:1742–1747
134. Friggens N, Van der Waaij L (2008) Modelling of resource allocation patterns. In: Rauw WM (ed) *Resource allocation theory applied to farm animal production*. CAB International, Wallingford, pp 302–320
135. Hughes BO, Duncan IJH (1988) The notion of ethological 'need', models of motivation and animal welfare. *Anim Behav* 36:1696–1707
136. Damm BI, Lisborg L, Vestergaard KS, Vanicek J (2003) Nest-building, behavioural disturbances and heart rate in farrowing sows kept in crates and Schmid pens. *Livest Prod Sci* 80:175–187
137. Ruis MAW, Te Brake JHA, Engel B, Buist WG, Blokhuis HJ, Koolhaas JM (2001) Adaptation to social isolation Acute and long-term stress responses of growing gilts with different coping characteristics. *Physiol Behav* 73:541–551
138. Held S, Cooper JJ, Mendl MT (2009) Advances in the study of cognition, behavioural priorities and emotions. In: Marchant-Forde JN (ed) *The welfare of pigs*. Springer, Berlin, pp 47–94
139. Manteuffel G (2002) Central nervous regulation of the hypothalamic-pituitary-adrenal axis and its impact on fertility, immunity, metabolism and animal welfare – a review. *Archive für Tierzucht* 45:575–595
140. Tuytens F (2007) Stereotypies. In: Velarde A, Geers R (eds) *On farm monitoring of pig welfare*. WAP, Wageningen, pp 41–46
141. Wemelsfelder F (2007) Apathy. In: Velarde A, Geers R (eds) *On farm monitoring of pig welfare*. WAP, Wageningen, pp 47–52
142. Mason G, Bateson M (2009) Motivation and the organization of behaviour. In: Jensen P (ed) *The ethology of domestic animals*, 2nd edn. CABI, Wallingford, pp 38–56
143. Beattie VE, Walker N, Sneddon IA (1995) Effects of environmental enrichment on behaviour and productivity of growing pigs. *Anim Welfare* 4:207–220
144. Beattie VE, O'Connell NE, Moss BW (1999) The influence of environmental enrichment on behaviour, performance and meat quality of domestic pigs. In: BSAS conference, Scarborough (Communication 192)
145. Webster J (2006) Ideals and realities: what do we owe to farm animals? In: Turner J, D'Silva J (eds) *Animals, ethics and trade*. Earthscan, London, pp 149–158
146. Jensen P (2009) Behaviour genetics, evolution and domestication. In: Jensen P (ed) *The ethology of domestic animals*, 2nd edn. CABI, Wallingford, pp 10–24
147. Jensen P (2002) Behaviour of pigs. In: Jensen P (ed) *The ethology of domestic animals*, 1st edn. CABI, Wallingford, pp 159–172
148. Stolba A, Wood-Gush DGM (1989) The behaviour of pigs in a semi-natural environment. *Anim Prod* 48:419–425
149. Kruska DCT (2005) On the evolutionary significance of encephalization in some eutherian mammals: effects of adaptive radiation, domestication, and feralization. *Brain Behav Evol* 65:73–108
150. Bolhuis JE (2004) Personalities in pigs: individual characteristics and coping with environmental challenges. Ph.D. thesis, Wageningen University, The Netherlands
151. Carroll BJ (2009) The neuroendocrinology of mood disorders. In: Pfaff DW, Arnold AP, Fahrbach SE, Etgen AM, Rubin RT (eds) *Hormones, brain and behavior*, vol 5, 2nd edn. Academic, San Diego, CA, pp 2899–2926
152. Loijens LWS (2002) Stress, endogenous opioids and stereotypies in tethered pigs. Ph.D. thesis, Wageningen University, The Netherlands
153. Karman AG (2003) Neuroendocrine adaptation to stress in pigs. Ph.D. thesis, Wageningen University, The Netherlands
154. Veenema AH (2003) Coping style and stressor susceptibility: neuroendocrine and neurochemical studies with genetically selected mouse lines. Ph.D. thesis, University of Groningen, The Netherlands
155. Benus RF, Bohus B, Koolhaas JM, Van Oortmerssen GA (1991) Heritable variation for aggression as a reflection of individual coping strategies. *Cell Mol Life Sci* 47:1008–1019
156. De Kloet ER (2004) Hormones and the stressed brain. *Ann NY Acad Sci* 1018:1–15
157. Keeling L, Jensen P (2009) Abnormal behaviour, stress and welfare. In: Jensen P (ed) *The ethology of domestic animals*, 2nd edn. CABI, Wallingford, pp 85–101
158. Ramos A, Mormède P (1998) Stress and emotionality: a multidimensional and genetic approach. *Neurosci Biobehav Rev* 22:33–57
159. Van Oortmerssen GA, Bakker TC (1981) Artificial selection for short and long attack latencies in wild *Mus musculus domesticus*. *Behav Genet* 11:115–126
160. Cools AR, Brachten R, Heeren D, Willemsen A, Ellenbroek B (1990) Search after neurobiological profile of individual-specific features of Wistar rats. *Brain Res Bull* 24:49–69
161. Touma C, Bunck M, Stein H, Zeh R, Landgraf R (2006) Mice selected for high or low stress reactivity: a new animal model for affective disorders. *Front Neuroendocrinol* 27:56
162. Morris R (2006) Stress and the hippocampus. In: Andersen P (ed) *The hippocampus book*. Oxford University Press, New York, pp 751–768
163. Poletto R, Steibel JP, Siegford JM, Zanella AJ (2005) Effects of early weaning and social isolation on the expression of glucocorticoid and mineralocorticoid receptor and 11 $\beta$ -hydroxysteroid dehydrogenase 1 and 2 mRNAs in the frontal cortex and hippocampus of piglets. *Brain Res* 1067:36–42
164. Kanitz E, Puppe B, Tuchscherer M, Heberer M, Viergutz T, Tuchscherer A (2009) A single exposure to social isolation in domestic piglets activates behavioural arousal, neuroendocrine stress hormones, and stress-related gene expression in the brain. *Physiol Behav* 98:176–185
165. Pottinger TG (2000) Genetic selection to reduce stress in animals. In: Moberg GP, Mench JA (eds) *The biology of animal*



- stress: basic principles and implications for animal welfare. CABl, Wallingford, pp 291–308
166. D'Eath RB, Conington J, Lawrence AB, Olsson IAS, Sandøe P (2010) Breeding for behavioural change in farm animals: practical, economic and ethical considerations. *Anim Welfare* 19(Suppl 1):17–27
  167. Mormède P (2005) Molecular genetics of behaviour: research strategies and perspectives for animal production. *Livest Prod Sci* 93:15–21
  168. Jensen P (1994) Fighting between unacquainted pigs. Effects of age and of individual reaction pattern. *Appl Anim Behav Sci* 41:37–52
  169. Couret D, Otten W, Puppe B, Prunier A, Merlot E (2009) Behavioural, endocrine and immune responses to repeated social stress in pregnant gilts. *Animal* 3:118–127
  170. Turner SP, Roehe R, D'Eath RB, Ison SH, Farish M, Jack MC, Lundeheim N, Rydhmer L, Lawrence AB (2009) Genetic validation of post-mixing skin injuries in pigs as an indicator of aggressiveness and the relationship with injuries under more stable social conditions. *J Anim Sci* 87:3076–3082
  171. Murani E, D'Eath RB, Turner SP, Evans G, Foury A, Kurt E, Thölking L, Klont R, Ponsuksili S, Mormède P, Wimmers K (2009) Identification of genes involved in the genetic control of aggressiveness, stress responsiveness, pork quality and their interactions. In: EAAP Barcelona, Spain (Communication 26-7)
  172. Terenina E, Bazovkina D, Rousseau S, Salin F, Monllor S, Kulikov A, Turner SP, D'Eath RB, Mormède P (2010) Association between aggressive behavior and candidate gene polymorphisms: study of the brain serotonergic system in pigs. In: 9th WCGALP, Leipzig, Germany (Communication 0864)
  173. Pauly C, Ampuero S, Bee G (2010) Expected effects on carcass and pork quality when surgical castration is omitted: results of a meta-analysis study. In: EAAP Heraklion, Greece (Communication 17-6)
  174. PigCas (2008) Attitudes, practices and state of the art regarding piglet castration in Europe. University of Newcastle, UK. [http://w3.rennes.inra.fr/pigcas/Public\\_reports/D3\\_3\\_Final\\_report\\_evaluation.pdf](http://w3.rennes.inra.fr/pigcas/Public_reports/D3_3_Final_report_evaluation.pdf)
  175. Squires EJ (2006) Possibilities for selection against boar taint. *Acta Vet Scand* 48(Suppl 1):S8
  176. Tajet H, Andresen Ø, Meuwissen THE (2006) Estimation of genetic parameters of boar taint; skatole and androstenone and their correlations with sexual maturation. *Acta Vet Scand* 48(Suppl 1):S9
  177. Merks JWM, Bergsma R, Bloemhof S, Roelofs-Prins DT, Knol EF (2010) Quantitative genetic opportunities to ban castration. In: EAAP Heraklion, Greece (Communication 17-7)
  178. Moe M, Lien S, Aasmundstad T, Meuwissen THE, Hansen MHS, Bendixen C, Grindflek E (2009) Association between SNPs within candidate genes and compounds related to boar taint and reproduction. *BMC Genet* 10:32
  179. Schenkel FS, Squires EJ (2010) Reducing boar taint in pigs using SNP markers. In: 9th WCGALP, Leipzig, Germany (Communication 0674)
  180. Breuer K, Sutcliffe MEM, Mercer JT, Rance KA, O'Connell NE, Sneddon IA, Edwards SA (2005) Heritability of clinical tail-biting and its relation to performance traits. *Livest Prod Sci* 93:87–94
  181. Turner SP, Roehe R, Lawrence AB (2010) Social behaviour in pigs. In: 9th WCGALP, Leipzig, Germany (Communication 0060)
  182. Taylor NR, Main DCJ, Mendl M, Edwards SA (2009) Tail-biting: a new perspective. *Vet J* 186(2):137–147
  183. Bolhuis JE, Schouten WGP, Schrama JW, Wiegant VM (2006) Effects of rearing and housing environment on behaviour and performance of pigs with different coping characteristics. *Appl Anim Behav Sci* 101:68–85
  184. Bracke MBM (2007) Vulva biting. In: Velarde A, Geers R (eds) On farm monitoring of pig welfare. Wageningen Academic Publishers, Wageningen, pp 65–70
  185. Harris MJ, Li YZ, Gonyou HW (2003) Savaging behaviour in gilts and sows. *Can J Anim Sci* 83:819–821
  186. Kjaer JB, Mench JA (2003) Behaviour problems associated with selection for increased production. In: Muir WM, Aggrey SE (eds) Poultry genetics, breeding and technology. CAB International, Wallingford, pp 67–82
  187. Su G, Kjaer JB, Sørensen P (2005) Variance components and selection response for feather-pecking behavior in laying hens. *Poult Sci* 84:14–21
  188. Labouriau R, Kjaer JB, Abreu GCG, Hedegaard J, Buitenhuis AJ (2009) Analysis of severe feather pecking behavior in a high feather pecking selection line. *Poult Sci* 88:2052–2062
  189. Wysocki M, Stratz P, Preuss S, Bennewitz J (2010) Functional investigation of candidate genes affecting feather pecking in chickens. In: 9th WCGALP, Leipzig, Germany (Communication 0254)
  190. Quilter CR, Gilbert CL, Oliver GL, Jafer O, Furlong RA, Blott SC, Wilson AE, Sargent CA, Mileham A, Affara NA (2008) Gene expression profiling in porcine maternal infanticide: a model for puerperal psychosis. *Am J Med Genet B Neuropsychiatr Genet* 147B:1126–1137
  191. Ellen ED, Visscher J, Van Arendonk JAM, Bijma P (2008) Survival of laying hens: genetic parameters for direct and associative effects in three purebred layer lines. *Poult Sci* 87:233–239
  192. Bergsma R, Kanis E, Knol EF, Bijma P (2008) The contribution of social effects to heritable variation in finishing traits of domestic pigs (*Sus scrofa*). *Genetics* 178:1559–1570
  193. Chen CY, Kachman SD, Johnson RK, Newman S, Van Vleck LD (2008) Estimation of genetic parameters for average daily gain using models with competition effects. *J Anim Sci* 86:2525–2530
  194. Muir WM, Craig JV (1998) Improving animal well-being through genetic selection. *Poult Sci* 77:1781–1788

195. Cheng H, Muir WM (2005) The effects of genetic selection for survivability and productivity on chicken physiological homeostasis. *Worlds Poult Sci J* 61:383–397
196. Cheng HW, Eicher SD, Chen Y, Singleton P, Muir WM (2001) Effect of genetic selection for group productivity and longevity on immunological and hematological parameters of chickens. *Poultry Sci* 80:1079–1086
197. Gunsett F (2005) Group selection in swine – a case study. National Swine Improvement Federation Conference, Ottawa, Canada. [www.nsisf.com/Conferences/2005/pdf/GroupSelectionSwine.pdf](http://www.nsisf.com/Conferences/2005/pdf/GroupSelectionSwine.pdf)
198. Muir WM, Schinckel AP (2002). Incorporation of competitive effects in breeding pro-grams to improve productivity and animal well being. In: 7th WCGALP, Montpellier, France (Communication 14-07)
199. Bijma P (2009) Maintaining fitness by within breed selection. In: Van der Werf JHJ, Graser HU, Frankham R, Gondro C (eds) *Adaptation and fitness in animal populations*. Springer, Berlin, pp 103–124
200. Conington J, Gibbons J, Haskell MJ, Bünger L (2010) The use of breeding to improve animal welfare. In: 9th WCGALP, Leipzig, Germany (Communication 0057)
201. Faure JM (1980) To adapt the environment to the bird or the bird to the environment? In: Moss R (ed) *The laying hen and its environment*. Nijhoff, Den Haag, pp 19–42
202. Garnett T (2009) Livestock-related greenhouse gas emissions: impacts and options for policy makers. *Environ Sci Policy* 12:491–503
203. Hörning B (2008) Auswirkungen der Zucht auf das Verhalten von Nutztieren. Tierzuchtfonds für artgemässe Tierzucht, Kassel University Press, Kassel, Germany
204. Lee DHK (2006) Problems in the environmental adaptation of domestic animals. *Ann NY Acad Sci* 91:608–616
205. François D, Boissy A, Jacquet P, Allain D, Bibé B, Rupp R, Moreno CR, Bodin L, Hazard D, Bouix J (2010) Genetics of adaptation traits for harsh environment in sheep. In: 9th WCGALP, Leipzig, Germany (Communication 0453)
206. Herre W, Röhrs M (1990) *Haustiere – zoologisch gesehen*, 2nd edn. Fischer Verlag, Stuttgart
207. Belyaev DK, Trut LN (1975) Some genetic and endocrine effects of selection for domestication in silver foxes. In: Fox MW (ed) *The wild canids*. Van Nostrand Reinhold, New York, pp 416–426
208. Reiss MJ, Straughan R (1996) *Improving nature? The science and ethics of genetic engineering*. Cambridge University Press, Cambridge
209. Webster J (2005) Ideals and realities: what do we owe to farm animals? In: *From Darwin to Dawkins: the science and implications of animal sentience (DVD)*. Compassion in World Farming Trust, Petersfield
210. FAWC (2004) Report on the welfare implications of animal breeding and breeding technologies in commercial agriculture. Farm Animal Welfare Council, London
211. Appleby MC, Sandøe PT (2002) Philosophical debate on the nature of well-being: implications for animal welfare. *Anim Welfare* 11:283–294
212. Thompson PB (2010) Why using genetics to address welfare may not be a good idea. *Poult Sci* 89:814–821
213. Rollin BE (2003) Ethics and species integrity. *Am J Bioeth* 3:15–17
214. Gavrell Ortiz SE (2004) Beyond welfare: animal integrity, animal dignity, and genetic engineering. *Ethics Environ* 9:94–120
215. McLnerney J (2004) *Animal welfare, economics and policy*. Defra, London. [www.defra.gov.uk/evidence/economics/foodfarm/reports/documents/animalwelfare.pdf](http://www.defra.gov.uk/evidence/economics/foodfarm/reports/documents/animalwelfare.pdf)
216. Würbel H (2009) Ethology applied to animal ethics. *Appl Anim Behav Sci* 118:118–127
217. De Boer IJM, Cornelissen AMG (2002) A method using sustainability indicators to compare conventional and animal-friendly egg production systems. *Poult Sci* 81: 173–181
218. Cederberg C, Flysjö A (2004) Environmental assessment of future pig farming systems: quantifications of three scenarios from the FOOD21 synthesis work. Swedish Institute for Food and Biotechnology, Göteborg, Sweden, SIK-report 723. [chaos.bibul.slu.se/sll/institutet\\_livsm\\_bioteknik/sik-rapport/SIK723/SIK723.pdf](http://chaos.bibul.slu.se/sll/institutet_livsm_bioteknik/sik-rapport/SIK723/SIK723.pdf)
219. Stern S, Sonesson U, Gunnarsson S, Kumm KI, Öborn I, Nybrant T (2005) Sustainable pig production in the future: development and evaluation of different scenarios. Swedish University of Agricultural Sciences, Uppsala, Sweden, Report FOOD21-5/2005. [citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.7315&rep=rep1&type=pdf](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.7315&rep=rep1&type=pdf)
220. Stott AW, Gunn GJ (2008) Use of a benefit function to assess the relative investment potential of alternative farm animal disease prevention strategies. *Prev Vet Med* 84:179–193
221. Lokhorst C, Groot Koerkamp PWG (2009) *Precision livestock farming '09*. Wageningen Academic Publishers, Wageningen
222. Krüger L (1961) Geschichtliche Entwicklung der Rassen in der europäischen Tierzucht. In: Hammond J, Johansson I, Haring F (eds) *Handbuch der Tierzüchtung 3-1: Rassenkunde*. Parey, Hamburg, pp 25–53
223. Mason IL (1969) *A world dictionary of livestock breeds, types and varieties*, 4th edn. CAB International, Wallingford
224. Porter V (1993) *Pigs. A handbook to the breeds of the world*. Helm Information, Mountfield, Robertsbridge
225. Falkenberg H, Hammer H (2007) Zur Geschichte und Kultur der Schweinezucht und -haltung. 3: Schweinezucht und -haltung in Deutschland von 1650 bis 1900. *Züchtungskunde* 79:92–110
226. Huisman A, Chereil P, Van Haandel B (2010) Linkage disequilibrium and signatures of selection on chromosome 1 in a commercial sire and dam line. In: 9th WCGALP, Leipzig, Germany (Communication 0840)

## Plant Breeding Under a Changing Climate

M. FERNANDA DRECCER<sup>1</sup>, DAVID BONNETT<sup>2</sup>,  
TANGUY LAFARGE<sup>3,4</sup>

<sup>1</sup>CSIRO Plant Industry Cooper Laboratory, Gatton,  
QLD, Australia

<sup>2</sup>CIMMYT Int, Mexico D.F, Mexico

<sup>3</sup>CIRAD, UMR AGAP, Montpellier, France

<sup>4</sup>IRRI, CESD, Los Baños, Philippines

### Article Outline

Glossary

Definition of the Subject

Introduction

Breeding in a Changing Environment

A Combination of Breeding Approaches Needed to  
Advance Yield in a Changing Climate

A Niche for Indirect Selection for Yield Using Physio-  
logical Parameters

Identification of Variation for Physiological Traits from  
Exotic Germplasm Sources

Conclusions and Future Directions

Acknowledgments

Bibliography

### Glossary

**Phenotyping** The activity of measuring the physiological, morphological, developmental, and chemical characteristics of plants.

**Trait** A measurable phenotypic character or attribute, for example, plant height.

### Definition of the Subject

The next generation of crops, capable of being productive in an increasingly variable and changing climate, will rely on genetic interventions based on process understanding, selection of target traits in managed environments, and high-throughput phenotyping and genotyping more than ever before. This entry discusses examples from wheat and rice, recent advances in plant breeding for high yield potential environments, and also those where abiotic stress is a major limitation to productivity. The

methodologies and lessons learnt are discussed in the context of breeding in the face of climate change.

### Introduction

The effects of climate change on agricultural production and food security are already taking place, creating new challenges for plant breeders to act quickly. The consequences of climate change on agricultural systems across the globe will be heterogeneous [35]. The projections for 2050 indicate that the increase in temperature (1–3°C) and CO<sub>2</sub> together with rainfall changes may benefit crops in the mid- to high latitudes, as temperatures will be closer to optimal for growth and the growing season longer. Over the same period, a decline in agricultural productivity is projected for low-latitude agricultural systems due to detrimental thermal conditions and more frequent extreme weather-related events. In the longer term, if the effects of climate change are not counteracted, productivity could decline both in low and mid- and high latitudes, primarily due to detrimental impacts of high temperatures and water stress [35, 66]. Rising temperatures will lower production by limiting the length of the growing season, exerting direct negative effects on resource capture and processes underpinning growth and yield. Another consequence of rising global temperatures over the next few decades is likely to be the increase in evaporation and acceleration of the global hydrological cycle, which could potentially dry subtropical areas and increase precipitation at higher latitudes. Ongoing challenges to food security will result from these changes, as most developing countries are situated at low latitudes in regions that are already warm and semi-arid [66]. To illustrate this point, two thirds of the undernourished people in the world live in just seven countries (Bangladesh, China, the Democratic Republic of the Congo, Ethiopia, India, Indonesia, and Pakistan) and over 40% live in China and India alone [23].

While general trends are described above, changes can already be observed. In Australia, average temperatures have increased 0.9°C since 1950, with significant regional variation, while the frequency of hot days and nights has increased and that of cold days and nights has declined ([www.climatechangeinaustralia.com.au](http://www.climatechangeinaustralia.com.au)). In parallel, since 1950, most of Eastern and Western Australia has experienced substantial rainfall decline,

while North-West Australia has become wetter. In this context, new crops and crop varieties represent a technical adaptation with the potential to be instrumental in at least reducing climate-related vulnerability at the farm level [33]. For Australia's wheat crops, it has been estimated that, in the absence of adaptive measures, a 1.5–2°C increase in temperature would cancel out the grain yield increase derived from a CO<sub>2</sub> doubling, assuming no change in varietal adaptation ([33] and references therein).

It is important to consider that plant breeding takes time. The objective of a plant breeding program is to create new genetic variation and select gene combinations to create genotypes with superior performance in the target population of environments (TPE) [16]. Combining a range of methods, from traditional plant breeding to molecular tools, it is estimated that it takes 7–12 years to release a wheat cultivar (David Bonnett, pers. comm.) and 5–10 years to release a rice cultivar [9]. Increased climate variability in terms of rainfall patterns and the trends in the evolution of major weather variables such as temperature will lead to longer-term changes in the TPE. Under increased weather variability, a higher genotype × environment interaction (GEI) is expected. An increase in GEI, observed through altered genotypic rankings, makes it harder for breeders to make sustained genetic gains, as already documented for drought [65]. The paradox is that at a time when farmers' needs for new varieties as an adaptation tool intensify, breeding progress may become slower.

In this context, three of the main challenges plant breeding faces in relation to adaptation to climate change are (1) identifying the new target population of environments (TPE), (2) translating this knowledge into practical selection methods to uncover new genetic variation, including large mass phenotyping of potential parental lines, progeny, and wild genetic resources, and (3) integrating complex genotypic information with the large volume of data from high-throughput phenotyping systems. This entry looks at some of the recent advances in plant breeding for high yield potential environments and also those where abiotic stress is a major limitation to productivity. The methodologies and lessons learnt could become useful when breeding in the face of climate change. Examples are given for rice and wheat, because of their important contribution in volume and value to the world economy [24].

## Breeding in a Changing Environment

Yield is a complex trait underpinned by many different processes and, as such, highly influenced by environmental conditions. Breeding programs utilize multi-environment trials as a way of sampling the target population of environments (TPE). However, the conditions (weather, soil, agronomy) in those trials are not always a good representation of the TPE that the lines will grow in during their commercial life ([13] and references therein). This gap, between the “selection” TPE and the “commercial life TPE” may increase as a result of increased climate variability. Predicting which type of environments breeders will be targeting and their frequency of occurrence may become a key piece of information in designing the best targeting of selection schemes.

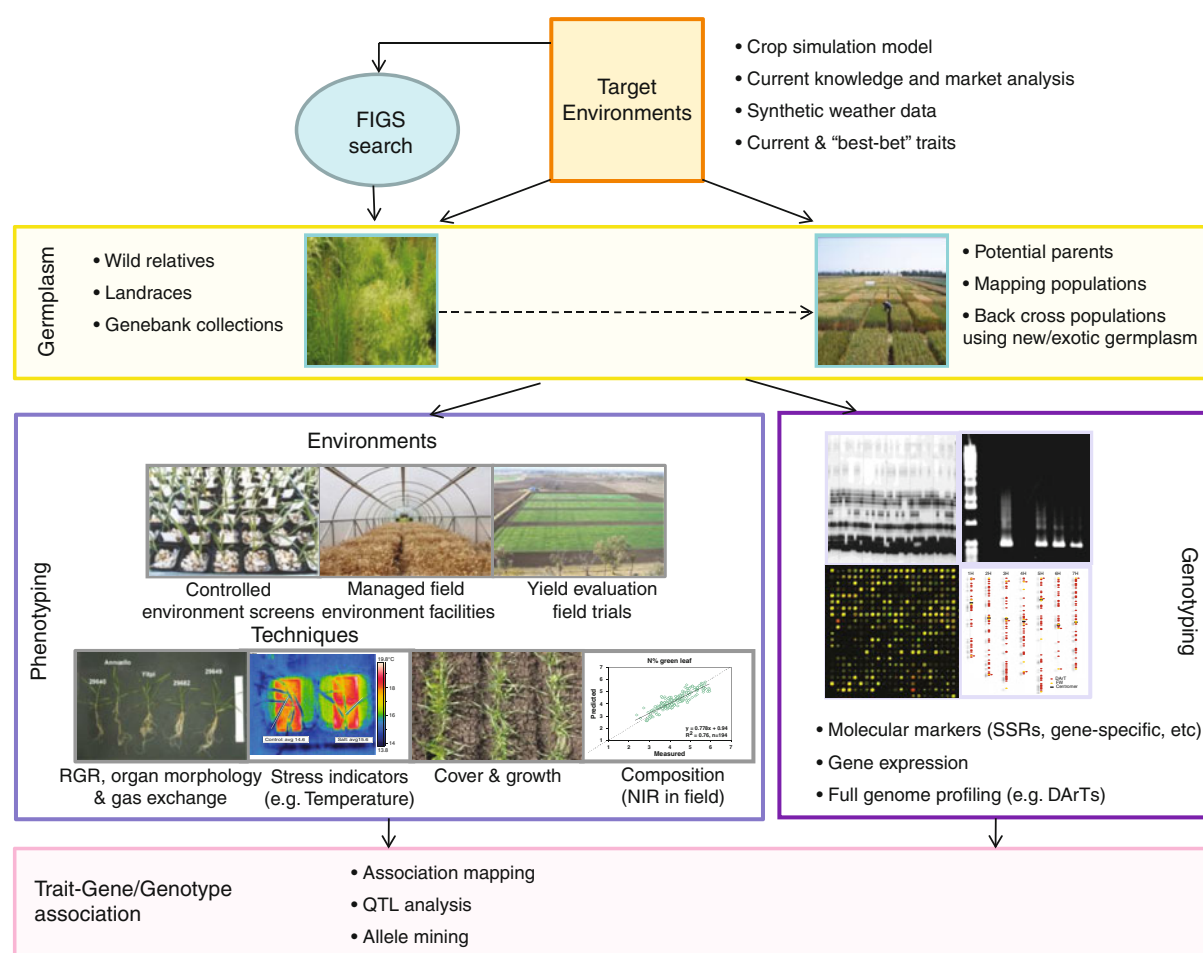
Some lessons can be learnt from experience in breeding for drought-prone environments that could be adopted more widely and potentially extended to selecting for adaptation to high temperatures. In drought-prone areas, an important proportion of the GEI has been attributed to the timing of drought stress with respect to the crop stage. Not surprisingly, different outcomes in yield progress can be expected when selection takes place under different drought patterns [13]. An attempt to describe “types of environment” for a particular region has been conducted for sorghum [13] and wheat [15] in Northern Australia utilizing historic weather data, soil characteristics, and current and virtual crop characteristics (e.g., sorghum: [14]) using the crop simulation model APSIM [38]. This could be expanded to a larger regional scale using synthetic weather data (“future climates”) aiming at different outcomes. Although climate projections have their own uncertainty, there is consensus in some of the predicted global trends [32]. At a large scale, describing “types of environments” using synthetic weather data could be a tool to identify shifts in the cultivated area and regions that are likely to experience increased frequency of events that will lead to yield loss [42], for example, high temperature or drought during pollen meiosis in wheat [37] or high temperatures at anthesis in rice [36]. This information can be used to design the layout of multi-environment trials or create managed environment facilities that reliably “reproduce” these conditions, for example, irrigation in the desert (such

as CIMMYT's Obregon facility), rain-out shelters, temperature gradient tunnels, etc., that, depending on their scale, could be used for particular stages of the breeding program. Ultimately, this information could be used to set selection criteria prioritizing particular traits (e.g., [67]) and devise the appropriate tests and technologies to screen for them (Fig. 1).

### A Combination of Breeding Approaches Needed to Advance Yield in a Changing Climate

A modern wheat variety carries an elite combination of alleles for productivity, grain quality, and resistance or

tolerance to a suite of the most important biotic and abiotic stresses in the environments in which they are grown. In a breeding program targeting development of new varieties for production by farmers, it is usually necessary to make crosses between parents with a high co-ancestry and many attributes in common while differing for a small number of traits important to farmers and end users. Often they will be current varieties or advanced breeding lines. With this approach, the gains made in one breeding cycle are built on in subsequent cycles and progress is incremental. However, in spite of this relative conservatism, most crosses still fail to produce their desired outcome, and breeders



**Plant Breeding Under a Changing Climate. Figure 1**

Example of how to target traits to introduce in a breeding program for a changing climate. Defining target environments can help direct a FIGS search for exotic germplasm and set up specific populations. Phenotyping will be necessary both in the exotic germplasm and the breeding populations. Genotyping information combined with phenotypic data is the basis for trait–gene association

must make multiple crosses to spread risk. In drier environments subject to greater GEI, progress for yield is usually slower than in more favorable environments. In spite of these constraints, breeding programs have successfully improved water-limited yields over an extended period and new approaches must offer tangible benefits if they are to be adopted [46, 52, 75].

The already commonplace and increasing use of molecular markers in wheat breeding is a good example of the adoption of a technology that successfully supplements conventional approaches. Most markers implemented so far have been for genes of large effect associated with disease resistance, grain quality, or, more recently, the major genes controlling plant phenology. Few if any “yield,” “drought tolerance” or “heat tolerance” quantitative trait loci (QTL) have been used in wheat breeding to date. Nonetheless, the major gene markers are useful in allowing more effective selection in earlier generations for a greater range of traits, while still leaving sufficient variation to allow selection of superior individuals for quantitative traits in later generations through conventional phenotypic selection (e.g., [77]). Removal of inferior individuals and alleles earlier in the breeding process means that fewer individuals entering the expensive later stages of yield and quality evaluation will be discarded because they lack some key major genes.

The identification of robust, yield-related QTL markers should allow early generation enrichment of the frequencies of these alleles in much the same way as for major genes. These QTL can be identified through a number of approaches. Initially, biparental mapping populations were used but these have two key problems. First, the population sizes and number of test environments required to accurately identify QTL for yield are large, even in the absence of GEI. Usually, suboptimal population sizes and numbers of test environments were used, leading to some QTL not being identified and the effects of others being overestimated [5]. Other problems with this approach are that only two parents are sampled per population and identified QTL are often population- or environment-specific (e.g., [60]). Further, parents commonly differed for major genes controlling height or phenology and these had by far the largest effects on the quantitative trait being measured making it difficult

to identify previously unknown sources of variation in the face of further-reduced effective population sizes [54].

Association mapping and whole genome prediction models (e.g., based on full genome profiling techniques such as DArTs [39]) have been proposed as alternative methods to identify and combine useful variation for a range of traits [10, 30, 31, 43]. The advantages of this approach are that alleles from a greater range of parents can be examined which may have greater relevance to breeding populations if the set is appropriately formulated. In the context of a breeding program aiming to make progress for yield, the most appropriate lines are likely to be breeding lines, and the phenotyping will comprise the routine yield evaluation trials. A key advantage is that the resources allocated to phenotyping are likely to be much greater than is available to any stand-alone QTL mapping project and the associations identified are more likely to be relevant to breeding populations as they were developed in breeding populations. Given the likelihood that the effects of QTLs for complex traits will change over time through fixation of important regions and differing interactions with new alleles at other loci, a continual reassessment of the value of QTLs in breeding populations may be needed in parallel with their use in selection [57]. This requires integration of good multi-environment yield data and efficient whole genome fingerprinting techniques that can be applied to the large numbers of lines making up the yield trials of commercial breeding programs. This approach depends on large amounts of resource for phenotyping, genotyping, and information management that are currently not available within the public sector. Further, association mapping approaches may not give good clues to the underlying mechanisms responsible for the yield effects of QTLs, but given that conventional breeding has achieved yield gains despite ignorance of contributory mechanism(s), this need not be an impediment to their use. Given that there is not a perfect correlation between any one physiological parameter and yield and interrelationships between physiological characters and their effects on yield are often not well understood, widespread mapping and use of QTL for physiological characters are not likely in the public or private sectors in the near future.

### A Niche for Indirect Selection for Yield Using Physiological Parameters

Given that all increases in yield must have a physiological basis, it should in theory be possible to identify and select for this variation. “Physiological” or “trait-based” breeding has a niche value as complementary to more conventional crossing and selection methods [63, 73]. For selection of physiological characters to be viable in a breeding program a demonstrated genetic correlation with yield is a prerequisite as is development of a cheap, high-throughput selection tool. The ultimate outcome of selection for yield-related physiological characters must be greater genetic gain per breeding cycle or per unit of investment. The fact that physiological measurements do not depend on knowledge of the number and location of QTLs segregating in any given population means they can be used across a greater number of populations. This also means they can be used to screen exotic germplasm, such as wild relatives, sources for variation that may be based on novel alleles and could be introgressed into breeding populations. High order or composite traits evaluated in the field have been particularly successful as targets of this approach, compared to traits evaluated at the cellular level, which are more prone to be subject to upscaling problems [72]. This is also of value given that agronomically important complex traits have not been particularly amenable to improvement using marker-assisted selection [31].

An analytic physiological approach is also likely to be useful to improve candidate traits for which genetic variation is not readily available, as could be the case in the response to high temperatures, and a targeted search and introduction strategy is needed. To illustrate the possibilities, an example of the use of physiological traits-based breeding to cope with limited water in wheat and improve yield under favorable conditions in rice is presented below.

#### Packaging Traits to Cope with Limited Water in Wheat and Links with Breeding for High Temperatures

A number of relationships between yield and physiological parameters have been identified in wheat and indirect selection methods for yield subsequently implemented. A good example resulted from the

discovery of a positive relationship between irrigated yields and stomatal conductance in a historical series of CIMMYT wheats [25]. Subsequent research showed that selection for higher stomatal conductance could be used in indirect selection for increased yield in irrigated conditions [17]. Selection for high stomatal conductance using canopy temperature as a surrogate is now a routine procedure in the rainfed wheat program at CIMMYT (Manes, personal communication 2010).

Discovery of the relationship between  $C^{12}/C^{13}$  carbon isotope discrimination (CID) and yield under drought is another example that grew from postulation of a relationship based on theoretical considerations, subsequent identification of variation in wheat germplasm, demonstrating a relationship with yield, development of a selection tool and germplasm, and ultimately in release of improved varieties Drysdale and Rees [55, 56, 65].

Later genetic dissection identified QTL for CID in several mapping populations that had not been specifically developed for mapping CID and in which the parents were mainly commercial varieties [60]. Although the parents of these populations did not have the most extreme CID levels, the populations showed the genetic complexity of the trait, that diversity for CID alleles was present in current varieties and that it was possible to recover transgressive segregants for CID from these populations as extreme as any identified in previous germplasm surveys. Therefore, with knowledge of the relationship between CID and yield under drought, availability of an appropriate selection screen should allow breeders to indirectly select for higher yield in drought-prone environments. Although use of mass spectrometry to determine carbon isotope composition is a relatively expensive procedure and has not been applied routinely by commercially focused breeding programs, it was successfully applied in germplasm development efforts at CSIRO that led to the release of varieties Drysdale and Rees in collaboration with varietal breeding programs [55, 56]. Development of cheaper techniques to screen for the increased WUE that result in the CID differences may be applicable on a more routine basis in breeding programs.

QTL mapping of water soluble carbohydrate (WSC) contents in stem [60] and coleoptile length [59]

reveal similarly widespread and potentially useful variation for these traits in elite germplasm. In cereals, water soluble carbohydrates (WSC) stored in stems have been acknowledged as contributing to maintenance of grain filling rate when photosynthesis declines due to various stresses, for example, drought ([6, 47–81], heat stress [7], and possibly disease [8]. Increased coleoptiles length can be a useful option for systems utilizing moisture-seeking strategies, such as sowing in deep furrows. This is likely to be a useful combination of traits for climate change in Australia, with the projected temperature increase and the decrease in average annual and winter rainfall (fewer and drier sowing opportunities) in the southern areas of the wheatbelt toward 2030 [32]. Progress for these traits could be made by selecting variation already present in breeders' populations if an efficient selection screen were available.

As mentioned before some of the traits targeted for drought stress are potentially useful under heat stress, which is particularly the case for those related to transpirational cooling [54]. Increased root growth in a soil profile with water available at depth can increase the transpirational cooling of the crop, uncoupling it from air temperature and helping keep tissues in a "safer" temperature window. Epicuticular waxiness is another trait with a dual function, reducing heat load and transpiration. Waxiness can be scored visually in the field, but, despite the theoretical impact there have been no comprehensive studies of its impact in crops [64]. Flowering time has been exploited under terminal drought as a simple way of manipulating the water balance, early flowering leaving more water available to be used during grain filling. Early flowering can also be used to avoid high temperatures during grain filling. In both cases, advancing flowering can carry the penalty of lower biomass at flowering and increased probability of frost damage.

Some processes are directly affected by high temperatures, among them respiration, inflorescence fertility, and starch composition, and hence grain quality ([2, 4, 78, 82]). Night respiration and photorespiration are processes directly affected by temperature; however, the lack of an easy way to phenotype large number of lines and study its effects at the crop level makes it inaccessible as a target for breeders at this point in time. Susceptibility of pollen to high temperatures

and traits contributing to heat tolerance (biochemical mechanisms) and avoidance (e.g., anther dehiscence early in the day) is a topic much researched in rice (see [78] and references therein). Pollen sterility induced by drought and genetic variation for it has been confirmed in wheat [37] and could also be potentially triggered by high temperatures. This is likely to be a trait to be screened for in controlled environment facilities or using molecular markers, given how unpredictable high temperatures can be at a particular crop stage in the field. For the purpose of marker development or QTL identification it will be important to "detangle" the phenotype appropriately as, for instance, reduced pollen sterility under high temperature could occur due to lower tissue temperature in a line that has high transpirational cooling due to higher stomatal conductance.

#### **Physiological Traits to Raise Yield Potential in Rice: Different Targets for Temperate Versus Tropical Regions**

As highlighted earlier, an increase in the length of the growing season as well as improved growing seasonal conditions are forecasted for high latitudes under climate change. This section illustrates the experience in rice in breeding for increased yield potential. Yield potential is defined as grain yield only limited by incoming radiation and temperature at a given site. Most of rice production is derived from tropical and subtropical areas, where it is grown with irrigation water from the monsoon [79]. The yield potential of rice in the tropics has been stable at 9–10 t ha<sup>-1</sup> for the last 20 years [51]. However, the arable land for rice is continuously decreasing as a consequence of increasing urbanization, in parallel with the increase in the population of rice consumers. An increase in yield potential of rice of 10–15% is now necessary to cope with this raising demand [69]. Among the main limitations to rice yield potential in the tropics are (1) the limited amount of incoming radiation (combination of short days and cloudiness), (2) high relative humidity underlying high resistance to transpiration, and (3) a trend toward a short crop cycle to allow the growth of two to three crops per year. Indeed, the highest rice yield potential today is slightly higher or similar to that of IR64, that is, comparable to introgression lines of IR64



background [29], a benchmark inbred line which was developed by IRRRI breeders in the middle of the 1980s. Most of the increases in rice yield in the field have been achieved with hybrid rice [50]. In contrast, the main gain in crop productivity in the last 20 years has been observed under nutrient or water limitation [34]. Grain quality has also been a great focus of the last 20 years and has diversified significantly to meet the variable expectations of consumers from different regions [27].

The International Rice Research Institute was a pioneer in using physiological concepts in breeding for an ideotype, as illustrated by the so-called “new plant type” which reached limited success [51]. These guiding principles are still utilized in current efforts to improve yield potential in rice focusing in increasing biomass and harvest index as discussed below. One challenging option is to develop a rice plant with the  $C_4$  photosynthesis pathway: the radiation use efficiency would be increased considerably and so the yield potential up to by 30–50% [44, 68, 80]. It is taking an integrated program involving molecular biologists, geneticists, biotechnologists, and physiologists working together for a considerable number of years. The search for genetic variation in different aspects of the  $C_4$  pathway, such as leaf anatomical and cellular specialization and variation in mechanisms underlying the  $CO_2$  compensation point is already presenting a considerable phenotyping challenge [83, 84]. Another option, still challenging but perhaps more realistic in a shorter timeframe, is to improve the current plant types for high yield. In the tropics, both improved biomass accumulation and partitioning underpin the superiority of hybrid rice versus rice elite inbred lines, with margins between 10% and 20% from wet to dry seasons [11, 40, 49, 50]. It is possible to assume that breeding programs for yield potential could gain much by incorporating traits relevant to hybrid rice superiority into improved inbred lines and lead to a substantial yield gain [41].

Work conducted at the International Rice Research Institute has extensively examined the basis for yield differences between hybrid rice and elite inbred lines of similar crop duration [11, 12, 40]. With an initial focus in the tropics, these authors confirmed, under a range of contrasting conditions, that superior biomass production in the succeeding phenological phases and improved partitioning play a significant role in the

higher yields of the hybrids. Hybrids are characterized by (1) higher crop growth rate during each phenological phase leading to overall higher plant biomass at maturity, (2) earlier cessation of tiller production (associated with earlier biomass partitioning to culm and earlier accumulation of reserves) with similar tiller production rate, (3) larger pool of reserves in the culm at anthesis (estimated through a lower value of specific culm length, SCL), (4) larger remobilization of the accumulated reserves from the culm to the panicle during grain filling (associated with quicker grain filling and higher SCL at maturity), and (5) lighter unfilled spikelets indicating that grain filling was more efficient with less partially filled spikelets (associated with larger number of filled grains) [40, 41]. It is clear that it is worth looking for genetic variation in storage of soluble sugars in the stem as well as remobilization capacity, in line with results in wheat [81] and in view of identifying the driving force of the dynamics of soluble sugars. Bueno et al. [12] speculated that improved partitioning, more than source supply, is the key component driving crop performance of high-yielding genotypes in the tropics where variability in biomass accumulation among genotypes is poorly expressed due to low evaporative demand. The sink strength index, as an improved harvest index taking into account the culm vigor [12, 41], can be used as an integrated trait for screening genotypes with high or low partitioning efficiency. It cannot, however, be considered as a “foundation” trait for high yield potential, unlike the traits cited earlier, and seems rather to be the integrated expression at maturity, as sink size at anthesis, of the cumulated higher efficiency of more simple traits. Some other important considerations concern remobilizing assimilates from senescing to productive tillers, avoiding lodging and delaying root senescence during grain filling. Maintaining functional roots throughout grain filling, and maintaining nitrogen uptake, should help delaying leaf senescence; however, leaf senescence has to be fast toward the end of grain filling to maximize remobilization.

### Identification of Variation for Physiological Traits from Exotic Germplasm Sources

While the wheat example cited above indicates considerable variation for several yield-related physiological

traits already existing in varieties and advanced breeding lines adapted to drier environments, for some traits, key genetic variation is lacking in existing breeding material. An example of this problem is the difficulty of recovering long coleoptile semidwarfs in wheats carrying the semidwarf alleles *Rht-B1b* and *Rht-D1b* [1, 76]. These alleles are virtually ubiquitous in modern semidwarf genes but have negative pleiotropic effects on coleoptile length. In order to develop semidwarf wheats with substantially longer coleoptiles the introduction of novel dwarfing genes that do not affect coleoptile length is necessary [22, 58]. For other traits such as root depth, likely to be related to yield under drought or heat stress, variation has been found in landraces and wild or synthetic hexaploid material that is greater or absent in existing wheat germplasm available to breeders [61, 62]. The introduction and exploitation of variation from synthetic hexaploid wheats in the CIMMYT wheat program is an example of the potential gains that can be made from exotic germplasm sources. These synthetic wheats were produced by re-synthesizing bread wheat from progenitor species the tetraploid durum wheat and the diploid wild grass *Triticum tauschii* [70, 71]. Derivatives of crosses between synthetic hexaploids and bread wheat now comprise around 30% of the breeding populations in CIMMYT's rainfed wheat program and the best lines have superior yield under drought stressed and more favorable environments than the best conventional wheats [18]. While this demonstrates the possibility of gains from exotic germplasm sources, a more targeted approach in which exotic sources are pre-screened for traits related to yield under drought or high temperatures may produce even greater gains. A possible pathway of integration is shown in Fig. 1. In many cases novel yield-related variation in exotic germplasm sources would be difficult or impossible to identify simply by screening directly for yield because it is present in agronomically poor backgrounds. In such instances, screening for physiological traits likely to be related to yield may be a useful precursor to crossing to introduce new variation [62].

A number of other approaches may allow selection of better initial material even prior to screening for physiological traits. Molecular diversity studies have shown, for example, that genetic diversity of emmer wheat is greater than that in durum wheat [19]

and studies of synthetic wheats produced by crossing emmers with *A. tauschii* have shown greater yield under drought stress in Mexico, Pakistan, and Eastern India than synthetics produced by durum wheat x *A. tauschii* crosses [71]. Given the distribution of emmer wheats in drought-prone Mediterranean environments the useful variation present in this material may have been predictable. Better predictions based on a greater array of climatic, soil, and location data as exemplified by the focused identification of germplasm strategy or FIGS (e.g., [20, 21]) should further improve the quality of germplasm selected to screen for variation in phenotypic traits (Fig. 1).

Robust and time-efficient phenotyping is also critical for trait-based selection of potential parents for crossing blocks and evaluation of the progeny, such as needed to underpin physiological breeding [63]. Non-invasive technologies, such as those based on spectral reflectance and thermal sensing have a role in the identification and selection of traits in a breeding context, by allowing several crop characteristics to be surveyed in a single measurement at the crop scale [45] (Fig. 1). For example, spectral reflectance has been used to simultaneously survey canopy cover, nitrogen, and water status of the crop [25, 48], while canopy temperature has been used as an indicator of leaf conductance and water use [85]. Potential for continuous developments for high throughput include introducing changes in platforms (wireless systems, unmanned aerial vehicles, etc.) to allow more frequent data capture and greater area coverage. While glasshouse or lab-based screens have been indicated as an option for some traits (e.g., [28, 74]), in most cases, there is a low correlation between this level of evaluation and field-based rankings. The interactions involved not only in upscaling to the crop/canopy level but with the changing environment itself get in the way (examples in [72]). Instead, field canopy scale measurements have more potential since selection can take place at the crop level, which avoids scaling up issues (e.g., [62]). For instance, reflectance-based indices have proved very useful as indirect selection criteria to increase the efficiency of selection in wheat growing in a reasonably stable environment where the terminal drought is managed [3], while indicators of canopy cover and canopy temperature were a good proxy for performance or QTL detection in a hot environment [54].

## Conclusions and Future Directions

Improved crop varieties will be a vital component of adaptation to climate change. Plant breeding will have to operate at a higher level of efficiency to make the necessary genetic progress to address current and projected food needs. In this context, the three main challenges plant breeding faces in relation to climate change are (1) identifying the new target population of environments, (2) translating this knowledge into practical selection methods to uncover new genetic variation, including mass phenotyping of potential parental lines, progeny, and wild genetic resources, and (3) linking genetic and phenotypic information. How to link and interpret the new and comprehensive information on genotypic characteristics and the large volume of data generated by high-throughput phenotyping platforms will be a critical step toward selecting the next generation of traits fit to less predictable environments.

## Acknowledgments

The authors thank Lynne McIntyre (CSIRO) and Andrzej Kilian (DARts Pty Ltd) for contributing illustrations on genotyping. FD acknowledges the financial support of the Department of Agriculture, Fisheries and Forestry, CSIRO and the Climate Adaptation Flagship.

## Bibliography

### Primary Literature

- Allan RE (1989) Agronomic comparisons between Rht1 and Rht2 semi-dwarf genes in winter wheat. *Crop Sci* 29:1103–1108
- Atkin OK, Tjoelker MG (2003) Thermal acclimation and the dynamic response of plant respiration to temperature. *Trends Plant Sci* 8:343–351
- Babar MA, van Ginkel M, Reynolds MP, Prasad B, Klatt AR (2007) Heritability, correlated response, and indirect selection involving spectral reflectance indices and grain yield in wheat. *Aust J Agric Res* 58:432–442
- Barnabas B, Jager K, Feher A (2008) The effect of drought and heat stress on reproductive processes in cereals. *Plant Cell Environ* 31:11–38
- Beavis WD (1998) QTL analyses: power, precision, and accuracy. In: Patterson AH (ed) *Molecular dissection of complex traits*. CRC Press, Boca Raton, pp 145–162
- Bidinger F, Musgrave RB, Fischer RA (1977) Contribution of stored pre-anthesis assimilates to grain yield in wheat and barley. *Nature* 270:431–433
- Blum A, Sinmena B, Mayer J, Golan G, Shpiler L (1994) Stem reserve mobilisation supports wheat-grain filling under heat stress. *Aust J Plant Physiol* 21:771–781
- Blum A (1998) Improving wheat grain filling under stress by stem reserve mobilisation. *Euphytica* 100:77–83
- Brar D, Virk P (2010) How a modern rice variety is bred. *Rice Today* 9(Jan–March):11–12
- Brescghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. *Genetics* 161(2):1165–1177
- Bueno CS, Lafarge T (2009) Higher crop performance of rice hybrids than elite inbreds in the tropics. 1. Hybrids accumulate more biomass during each phenological phase. *Field Crops Res* 112:229–237
- Bueno CS, Pasuquin E, Tubana B, Lafarge T (2010) Improving sink regulation, and searching for promising traits associated with hybrids, as a key avenue to increase potential of the rice crop in the tropics. *Field Crops Res* 118:199–207
- Chapman SC (2008) Use of crop models to understand genotype by environment interactions for drought in real-world and simulated plant breeding trials. *Euphytica* 161:191–208
- Chapman SC, Cooper M, Podlich D, Hammer GL (2003) Evaluating plant breeding strategies by simulating gene action and dryland environments. *Agron J* 95:99–113
- Chenu K, Cooper M, Hammer GL, Mathews KL, Dreccer MF, Chapman SC (2011) Environment characterisation as an aid to wheat improvement – interpreting genotype-environment interaction by modelling water-deficit patterns in north eastern Australia. *J Exp Bot* 62:1743–1755
- Comstock RE (1977) Quantitative genetics and the design of breeding programs. In: *Proceedings of the international conference on quantitative genetics*. Iowa State University Press, Ames, 16–21 Aug 1976, pp 705–718
- Condon AG, Reynolds MP, Rebetzke GJ, Ginkel M van, Richards RA, Farquhar GD (2007) Using stomatal aperture-related traits to select for high yield potential in bread wheat. In: *Wheat production in stressed environments. Proceedings of the 7th international wheat conference, Mar del Plata, 27 Nov–2 Dec 2005*, pp 617–624
- Dreccer MF, Chapman SC, Ogbonnaya FC, Borgognone MG, Trethowan RM (2008) Crop and environmental attributes underpinning genotype by environment interaction in synthetic bread wheats evaluated in Mexico and Australia. *Aust J Agric Res* 59:447–460
- Dreisigacker S, Kishii M, Lage J, Warburton M, Ogbonnaya FC, van Ginkel M, Brettell R (2008) Use of synthetic hexaploid wheat to increase diversity for CIMMYT bread wheat improvement. *Aust J Agric Res* 59(5):413–420
- El-Bouhssini M, Street K, Joubi A, Ibrahim Z, Rihawi F (2009) Sources of wheat resistance to Sunn pest, *Eurygaster integriceps* Puton, in Syria. *Genet Resour Crop Evol* 56:1065–1069
- El-Bouhssini M, Street K, Amri A, Mackay M, Ogbonnaya FC, Omran A, Abdalla O, Baum M, Dabbous A, Rihawi F (2011) Sources of resistance in bread wheat to Russian wheat

- aphid (*Diuraphis noxia*) in Syria identified using the Focused Identification of Germplasm Strategy (FIGS). *Plant Breeding* 130:96–97
22. Ellis MH, Rebetzke GJ, Chandler P, Bonnett DG, Spielmeier W, Richards RA (2004) The effect of different height reducing genes on early growth characteristics of wheat. *Funct Plant Biol* 31:583–589
  23. FAO (2010) The state of food insecurity in the world. Chief publishing policy and support branch office of knowledge exchange, research and extension FAO, Rome, p57
  24. FAO (2010) FAO Statistical yearbook 2010. Economic and social development department, FAO, Rome
  25. Fischer RA, Rees D, Sayre KD, Lu Z-M, Condon AG, Saavedra AL (1998) Wheat yield progress associated with higher stomatal conductance and photosynthetic rate and cooler canopies. *Crop Sci* 38:1467–1475
  26. Fitzgerald GJ, Rodriguez D, Christensen LK, Belford R, Sadras VO, Clarke TR (2006) Spectral and thermal sensing for nitrogen and water status in rainfed and irrigated wheat environments. *Precision Agric* 7:233–248
  27. Fitzgerald MA, McCough SR, Hall RD (2009) Not just a grain of rice: the quest for quality. *Trends Plant Sci* 14:133–139
  28. Gregory PJ, Bengough AG, Grinev D, Schmidt S, Thomas WTB, Wojciechowski T, Young IM (2009) Root phenomics of crops: opportunities and challenges. *Funct Plant Biol* 36:922–929
  29. Guan YS, Serraj R, Liu SH, Xu JL, Ali J, Wang WS, Venus E, Zhu LH, Li ZK (2010) Simultaneously improving yield under drought stress and non-stress conditions: a case study of rice (*Oryza sativa* L.). *J Exp Bot* 61:4145–4156
  30. Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. *Crop Sci* 49:1–12
  31. Heffner EL, Lorenz AJ, Jannink J-L, Sorrells ME (2010) Plant breeding with genomic selection: gain per unit time and cost. *Crop Sci* 50:1681–1690
  32. Hennessy KJ, Whetton PH, Preston B et al (2010) Climate projections. In: Stokes C, Howden M (eds) *Adapting agriculture to climate change-preparing Australian agriculture, forestry and fisheries for the future*. CSIRO Publishing, Collingwood, pp 21–48
  33. Howden SM, Gifford RG, Meinke H (2010) Grains. In: Stokes C, Howden M (eds) *Adapting agriculture to climate change-preparing Australian agriculture, forestry and fisheries for the future*. CSIRO Publishing, Collingwood, pp 21–48
  34. Horie T, Shiraiwa T, Homma K, Katsura K, Maeda S, Yoshida H (2005) Can yields of low land rice resume the increases that they showed in the 1980s? *Plant Prod Sci* 8:259–274
  35. IPCC (2007) *Climate change 2007: impacts, adaptation, and vulnerability. Contribution of working group II to the fourth assessment report of the intergovernmental panel on climate change*. Cambridge University Press, Cambridge, UK
  36. Jagadish SVK, Craufurd PQ, Wheeler TR (2007) High temperature stress and spikelet fertility in rice (*Oryza sativa* L.). *J Exp Bot* 58:1627–1635
  37. Ji XM, Shiran B, Wan JL, Lewis DC, Jenkins CLD, Condon AG, Richards RA, Dolferus R (2010) Importance of pre-anthesis anther sink strength for maintenance of grain number during reproductive stage water stress in wheat. *Plant Cell Environ* 33:926–942
  38. Keating BA, Carberry PS, Hammer GL, Probert ME, Robertson MJ, Holzworth D, Huth NI, Hargreaves JNG, Meinke H, Hochman Z, McLean G, Verburg K, Snow V, Dimes JP, Silburn M, Wang E, Brown S, Bristow KL, Asseng S, Chapman S, McCown RL, Freebairn DM, Smith CJ (2003) An overview of APSIM, a model designed for farming systems simulation. *Eur J Agron* 18:267–288
  39. Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Res* 29:e25
  40. Lafarge T, Bueno CS (2009) Higher crop performance of rice hybrids than elite inbreds in the tropics. 2. Does sink regulation, rather than sink size, play a major role? *Field Crops Res* 114:434–440
  41. Lafarge T, Bueno C, Pasuquin E, Wiangsamut B (2009) Biomass accumulation and sink regulation in hybrid rice: consequences for breeding programs and crop management. In: Xie F, Hardy B (eds) *Accelerating hybrid rice development*. International Rice Research Institute, Los Baños, p 698. Invited oral presentation, international symposium on hybrid rice, Changsha, Hunan, 11–15 Sept 2008, pp 453–474
  42. Li Y, Ye W, Wang M, Yan X (2009) Climate change and drought: a risk assessment of crop-yield impacts. *Clim Res* 39:31–46
  43. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
  44. Mitchell PL, Sheehy JE (2000) Performance of a potential C4 rice: overview from quantum yield to grain yield. *Stud Plant Sci* 7:145–163
  45. Montes M, Melchinger AE, Reif JC (2007) Novel throughput phenotyping platforms in plant genetic studies. *Trends Plant Sci* 12:433–436
  46. O'Brien L (1982) Victorian wheat yield trends, 1898–1977. *J Aust Inst Agric Sci* 48:162–168
  47. Palta JA, Kobata T, Turner NC (1994) Carbon and nitrogen in wheat as influenced by postanthesis water deficits. *Crop Sci* 34:118–124
  48. Peñuelas J, Fillela Y (1998) Visible and near-infrared reflectance techniques for diagnosing plant physiological status. *Trends Plant Sci* 3:151–156
  49. Peng S, Cassman KG, Virmani SS, Sheehy J, Khush GC (1999) Yield potential trends of tropical rice since the release of IR8 and the challenge of increasing rice yield potential. *Crop Sci* 39:1552–1559
  50. Peng S, Yang J, Laza MRC, Sanico A, Vesperas RM, Son TT (2003) Physiological bases of heterosis and crop management strategies for hybrid rice in the tropics. In: Virmani SS, Mao CX, Hardy B (eds) *Hybrid rice for food security, poverty alleviation, and environmental protection. Proceedings of the 4<sup>th</sup> international symposium on hybrid rice, 14–17 May 2002, Hanoi. IRRI, Los Baños, pp 153–170*

51. Peng S, Kush GS, Virk P, Tang Q, Zhu Y (2008) Progress in ideotype breeding to increase rice wheat consortium. *Field Crops Res* 108:32–38
52. Perry MW, D'Antuono MF (1989) Yield improvement and associated characteristics of some Australian spring wheat cultivars introduced between 1860 and 1982. *Aust J Agric Res* 40:457–472
53. Pinter PJ, Hatfield JL, Schepers JS, Barnes EM, Moran MS, Daughtry CST, Upchurch DR (2003) Remote sensing for crop management. *Photogram Eng Rem S* 69:647–664
54. Pinto RS, Reynolds MP, Mathews KL, McIntyre CL, Olivares-Villegas JJ, Chapman SC (2010) Heat and drought adaptive QTL in a wheat population designed to minimize confounding agronomic effects. *Theor Appl Genet* 121:1001–1021
55. *Plant Varieties Journal* (2002) Official Journal of Plant Breeder's Rights Australia, 15(1):74
56. *Plant Varieties Journal* (2004) Official Journal of Plant Breeder's Rights Australia, 17(2):254–261
57. Podlich DW, Winkler CR, Cooper M (2004) Mapping as you go: an effective approach for marker-assisted selection of complex traits. *Crop Sci* 44:1560–1571
58. Rebetzke GJ, Richards RA, Fischer VM, Mickelson BJ (1999) Breeding long coleoptile, reduced height wheats. *Euphytica* 106:159–168
59. Rebetzke GJ, Ellis MH, Bonnett DG, Richards RA (2007) Molecular mapping of genes for coleoptile growth in bread wheat. *Theor Appl Genet* 114:1173–1183
60. Rebetzke GJ, van Herwaarden AF, Jenkins C, Weiss M, Lewis D, Ruuska S, Tabe L, Fettel N, Richards RA (2008) Quantitative trait loci for soluble stem carbohydrate production in wheat. *Aust J Agric Res* 59:891–905
61. Reynolds MP, Mujeeb-Kazi A, Sawkins M (2005) Prospects for utilising plant-adaptive mechanisms to improve wheat and other crops in drought & salinity-prone environments. *Ann Appl Biol* 146:239
62. Reynolds MP, Dreccer MF, Trethowan R (2007) Drought adaptive traits derived from wheat wild relatives and landraces. *J Exp Bot* 58:177–186
63. Reynolds MP, Manes Y, Izanloo A, Langridge P (2009) Phenotyping for physiological breeding and gene discovery in wheat. *Ann Appl Biol* 155:309–20
64. Reynolds MP, Hays D, Chapman S (2010) Breeding for adaptation to heat and drought stress. In: Reynolds MP (ed) *Climate change and crop production*. CAB International, Wallingford, pp 71–91
65. Richards RA, Rebetzke GJ, Condon AG, van Herwaarden AF (2002) Breeding opportunities for increasing the efficiency of water use and crop yield in temperate cereals. *Crop Sci* 42:111–121
66. Rosenzweig C (2009) Climate change and agriculture. In: Meyers RA (ed) *Encyclopedia of complexity and systems science*, pp 1071–1082. doi:10.1007/978-0-387-30440-3\_70, Part 3
67. Semenov MA, Halford NG (2009) Identifying target traits and molecular mechanisms for wheat breeding under a changing climate. *J Exp Bot* 60:2791–2804
68. Sheehy JE, Ferrer AB, Mitchell PL, Elmido-Mabilangan A, Pablico P, Dionora MJA (2008) How the rice crop works and why it needs a new engine. In: Sheehy JE, Mitchell PL, Hardy B (eds) *Charting new pathways to C4 rice*. Proceedings of the international workshop, IRRI, Los Baños, pp 3–26
69. Swaminathan MS (2007) Science and shaping the future of rice. In: Aggarwal PK, Ladha JK, Singh RK, Devakumar C, Hardy B (eds) *Science, technology, and trade for peace and prosperity*. Proceedings of the 26<sup>th</sup> international rice research conference, 9–12 Oct 2006, New Delhi, pp 3–14
70. Trethowan RM, Reynolds MP, Sayre KD, Ortiz-Monasterio I (2005) Adapting wheat cultivars to resource conserving farming practices and human nutritional needs. *Ann Appl Biol* 146:404–413
71. Trethowan RM, Mujeeb-Kazi A (2008) Novel germplasm resources for improving environmental stress tolerance of hexaploid wheat. *Crop Sci* 48:1255–1265
72. Sinclair TR, Purcell LC, Sneller CH (2004) Crop transformation and the challenge to increase yield potential. *Trends Plant Sci* 9:70–75
73. Sinclair TR, Purcell LC (2005) Is a physiological perspective relevant in a 'genocentric' age? *J Exp Bot* 56:2777–2782
74. Sirault XRR, James RA, Furbank RTA (2009) A new screening method for osmotic component of salinity tolerance in cereals using infrared thermography. *Funct Plant Biol* 36:970–977
75. Vandeleur RK, Gill GS (2004) The impact of plant breeding on the grain yield and competitive ability of wheat in Australia. *Aust J Agric Res* 55:855–861
76. Whan BR (1976) The association between coleoptile length and culm length in semi-dwarf and standard wheats. *J Aust Inst Agric Sci* 42:194–196
77. Wang J, Chapman SC, Bonnett DG, Rebetzke GJ (2009) Simultaneous selection of major and minor genes: use of QTL to increase selection efficiency of coleoptile length of wheat (*Triticum aestivum* L.). *Theor Appl Genet* 119:65–74
78. Wassmann R, Jagadish SVK, Heuer S, Ismail A, Redona E, Serraj R, Singh RK, Howell G, Pathak H, Sumfleth K (2009) Climate change affecting rice production: the physiological and agronomic basis for possible adaptation strategies. *Adv Agron* 101:59–122
79. Wassmann R, Jagadish SVK, Sumfleth K, Pathak H, Howell G, Ismail A, Serraj R, Redona E, Singh RK, Heuer S (2009) Regional vulnerability of climate change impacts on Asian rice production and scope for adaptation. *Adv Agron* 102:91–133
80. Yin X, Struik PC (2009) C3 and C4 photosynthesis models: an overview from the perspective of crop modelling. *NJAS Wageningen J Life Sci* 57:27–38
81. Dreccer MF, van Herwaarden AF, Chapman SC (2009) Grain number and grain weight in wheat lines contrasting for stems soluble carbohydrate concentration. *Field Crops Res* 112: 43–54
82. Wardlaw IF, Wrigley CW (1994) Heat tolerance in temperate cereals—an overview. *Aust J Plant Phys* 21:695–703
83. Hibberd JM, Sheehy JE, Langdale JA (2008) Using C-4 photosynthesis to increase the yield of rice - rationale and feasibility. *Current opinion Plant Biol* 11:228–231

84. Furbank RT, von Caemmerer S, Sheehy JE, Edwards G (2009) C(4) rice: a challenge for plant phenomics. *Funct Plant Biol* 36:970–977
85. Pinter PJ, Jr, Zipoli G, Reginato RJ, Jackson RD, Idso SB, Hohman JP (1990) Canopy temperature as an indicator of differential water use and yield performance among wheat cultivars. *Agric Water Management* 18:35–48

## Plant Molecular Pharming, Industrial Enzymes

SAIFULLAH KHAN<sup>1</sup>, VIDYA RAJAN<sup>2</sup>, JOHN HOWARD<sup>2</sup>

<sup>1</sup>Plant Biotechnology Section, International Center for Chemical and Biological Sciences, HEJ research Institute, University of Karachi, Karachi, Pakistan

<sup>2</sup>Applied Biotechnology Institute, San Luis Obispo, CA, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Historical Background of Industrial Enzymes

Production of Recombinant Proteins in Plants

Regulation of Growth and Use of Transgenic Plants

Plants as Sustainable Sources of Industrial Enzymes

Future Directions

Bibliography

### Glossary

**Diafiltration** Method of cross-flow filtration that separates filtrate from solids.

**Industrial enzymes** Proteins that are used in commercial applications where very specific catalysts are needed.

**Pharmacognosy** The study of medicines derived from natural sources.

**Plant molecular pharming** Production of pharmaceuticals or industrial enzymes from genetically engineered plants.

**Sustainability** Production to meet present needs without compromising the ability of future generations to meet their own needs.

**Transgenic plants** Plants genetically engineered by the introduction of foreign genes using recombinant DNA technology.

### Definition of the Subject

Plants have been domesticated since around 10,000 years ago in the fertile Babylonian crescent [1] and husbandry and breeding techniques have been applied to increase yield and storage and to retard spoilage [2]. Plants have been used since time immemorial for their medicinal properties, and in ACE 78 Discorides first described 600 medicinal plants in *De Materia Medica*. Although the first synthetic drug, salicylic acid, entered the market in 1897, plants are still used for pharmacognosy – the preparation of drugs from natural sources [3]. The types of plants used for this purpose, however, are usually distinct from those for food, feed, or fiber. With the advent of molecular techniques, all plants now have the potential to serve as production vehicles for natural or engineered products that were previously limited to other hosts [4, 5]. Plant molecular pharming of industrial proteins refers to recombinant proteins used in industrial processes and produced in plants. Enormous quantities of a variety of enzymes go into the making of products such as paper, leather, detergents, pharmaceuticals, food, beverages, chemicals, and fabric, to name a few, and the economical production of these industrially important enzymes is crucial to commerce. This production must be balanced with the need for sustainability and environmental stewardship.

Sustainable production of industrial enzymes requires that resources are not over-exploited and that the environment is not polluted by wastes. The use of plants as “green” factories can meet both criteria. Plants are a renewable resource and thus generally are not over-exploited and wastes are biodegradable. The problems associated with fertilization runoff, spread of transgenic genes to non-target plants, and crop and land usage must be addressed to allay public concern about the use of transgenic plants. But, combined with modern farming and containment methods, transgenic plants have the potential to produce large quantities of target material safely and sustainably [6, 7].

### Introduction

This article provides a review of plant-based production of industrial enzymes as a sustainable solution to increasing demand. The history of interest in sustainable development, industry and industrial enzymes will

set the background for a detailed analysis of how plant-based production systems compare in terms of sustainability to processes that currently exist. A compilation of potentially useful industrial enzymes produced in plants is given and a discussion of the economic, social, and environmental sustainability advantages provided by plant-based production systems is also provided.

### Historical Background of Industrial Enzymes

In view of the importance of enzymes to industrial processes, a brief overview of the events leading to this relationship is provided below.

### Development of Applications for Industrial Enzymes

Until mechanization and the industrial revolution, most people eked out a subsistence living using human and animal power for agriculture, production, and transport of people and goods. The industrial revolution eventually made possible tremendous increases in commerce, and consequent social upliftment through improved earnings, literacy, and working conditions. There were costs, however, in degradation of the environment through unregulated expansion of industries and consequent pollution. In 1956, the British Parliament enacted the world's first clean air act. Since then, governments have imposed restrictions on locations of industries and the disposal of pollutants and effluents. These measures have helped considerably to alleviate the impact of pollutants on the environment and health [8], but have been costly to the industries themselves. For example, costs imposed on the meat and meat by-products industries, which include industrial enzymes, are passed on to the consumer through increased prices [9] which is often negative for commerce. Thus, the pressure mounts to lower prices across the board.

An assessment of the true cost of product manufacturing is provided by Life Cycle Assessments (LCA) which measures, usually by mass loadings, the cradle-to-grave impact of every stage from resource procurement, inputs for manufacture, and outputs of wastes throughout the process [10]. Although it is often difficult to predict the LCA of a process, especially impacts on biodiversity and environment, key parameters affecting LCA are renewability of the raw material resource and energy inputs. Enzymes lower the

activation energy of chemical reactions and are sourced from renewable resources, and thus should lower impact. A theoretical life cycle comparison between the production of biodiesel using inorganic and enzyme catalysis favored the latter as having lower environmental impact, lowered toxicity, and lowered greenhouse gas emissions, all attributed to lower steam heating requirements [11]. Production of large quantities of enzymes therefore comprises a key consideration for future industrial manufacturing processes.

### Current Uses of Industrial Enzymes

Today, many industries use enzymes to manufacture a variety of goods (see Table 1) from food to paper to high-value pharmaceuticals. The utilization of industrial enzymes has now extended to almost all industries handling organic compounds. Enzymes are used in the production of detergents, reagents for the analysis of drugs or blood components, food or food additives, for fiber processing or pulp processing in the paper industry, and for environmental purification. The method of enzyme use also varies, for example, as an enzyme preparation, on the surface of an insoluble carrier in a bioreactor, or a biosensor with the enzyme integrated into an electrode. The annual world industrial enzyme market (excluding pharmaceuticals) is in billions of dollars and is composed largely of enzymes used as detergent ingredients and for food processing applications [12, 13]. Thus, enzymes constitute a key lubricant of commercial success in many fields.

Enzymes are formally classified and given an Enzyme Classification (EC) number by the International Union of Biochemistry and Molecular Biology (IUBMB) based on the reaction that they catalyze (<http://www.chem.qmul.ac.uk/iubmb/>). Each major category is further divided into subclasses and sub-subclasses. Several enzymes with different trivial or common names may share the same EC number based on the type of reaction that they catalyze (see Table 2). For example, lipases and amylases share the same major category, hydrolases, because they catalyze the breakdown of substrate by hydrolysis. However, enzymes of industrial importance are generally referred to by their trivial name, and trivial names will be used in this review. The source of industrial enzymes can be fungal, bacterial, animal, or plant and with the advent

**Plant Molecular Pharming, Industrial Enzymes. Table 1** Enzymes used in various industrial segments and their applications (Reprinted from [13]. With permission from Elsevier)

Industry	Industrial enzyme	Industrial effect/application
Detergent laundry and dish wash	Protease	Protein stain removal
	Amylase	Starch stain removal
	Lipase	Lipid stain removal
	Cellulase	Cleaning, color clarification, anti-redeposition (cotton)
	Mannanase	Mannanan stain removal (reappearing stains)
Starch and fuel	Amylase,	Starch liquefaction and saccharification
	Amyloglucosidase	Saccharification
	Pullulanase	Saccharification
	Glucose isomerase	Glucose to fructose conversion
	Cyclodextrin-glycosyltransferase	Cyclodextrin production
	Xylanase	Viscosity reduction (fuel and starch)
	Protease	Protease (yeast nutrition-fuel)
Food including dairy	Protease	Milk clotting, infant formulas (low allergenic), flavor
	Lipase	Cheese flavor
	Lactase	Lactose removal (milk)
	Pectin methyl esterase	Firming fruit-based products
	Pectinase	Fruit-based products
	Transglutaminase	Modify visco-elastic properties
Baking	Amylase	Bread softness and volume, flour adjustment
	Xylanase	Dough conditioning
	Lipase	Dough stability and conditioning (in situ emulsifier)
	Phospholipase	Dough stability and conditioning (in situ emulsifier)
	Glucose oxidase	Dough strengthening
	Lipoxygenase	Dough strengthening, bread whitening
	Protease	Biscuits, cookies
	Transglutaminase	Laminated dough strengths
Animal feed	Phytase	Phytate digestability-phosphorus release
	Xylanase	Digestibility
	$\beta$ -glucanase	Digestibility
Beverage	Pectinase	De-pectinization, mashing
	Amylase	Juice treatment, low calorie beer
	$\beta$ -glucanase	Mashing
	Acetolactate decarboxylase	Maturation (beer)
	Laccase	Clarification (juice), flavor (beer), cork stopper treatment



Plant Molecular Pharming, Industrial Enzymes. Table 1 (Continued)

Industry	Industrial enzyme	Industrial effect/application
Textile	Cellulase	Denim finishing, cotton softening
	Amylase	De-sizing
	Pectate lyase	Scouring
	Catalase	Bleach termination
	Laccase	Bleaching
	Peroxidase	Excess dye removal
Pulp and paper	Lipase	Pitch control, contaminant control
	Protease	Biofilm removal
	Amylase	Starch coating, de-inking, drainage improvement
	Xylanase	Bleach boosting
	Cellulase	De-inking, drainage improvement, fiber modification
Fats and oil	Lipase,	Transesterification
	Phospholipase	De-gumming, lyso-lecithin production
Organic synthesis	Lipase	Resolution of chiral alcohol and amides
	Acyase	Synthesis of semisynthetic penicillin
	Nitrilase	Synthesis of enantiopure carboxylic acids
Leather	Protease	Unhairing, bating
	Lipase	De-pickling
Personal care	Amyloglucosidase	Antimicrobial (combined with glucose oxidase)
	Glucose oxidase	Bleaching, antimicrobial
	Peroxidase	Antimicrobial

of molecular techniques, many of the genes for these enzymes have been cloned and transformed into organisms that are easy and convenient to grow in an industrial setting. Optimizing gene expression and culture conditions can increase quantities of enzymes produced and consequently lower costs.

In medicine, the prohibitive cost of insulin isolated from human cadavers created a market for porcine insulin despite the negative side effects. Eli Lilly introduced recombinant insulin in 1982, which decreased the cost to a more manageable level through the use of higher-affinity analogs [14], thereby saving innumerable lives. This price decrease is attributed to the use of transgenic bacteria, improved production methods, and also to lower costs of high-purity protease used to cleave the insulin molecule to active form [15].

Another prominent example today is the use of enzymes in the bioconversion of grain to ethanol [16]. In this case the enzymes very efficiently break down starch to fermentable sugars that can then be used to make ethanol. In the near future the hope is that enzymes will also be used to convert cellulosic material into fermentable sugars as well.

The global industrial enzyme market increased from US \$1 billion in 1995 to \$1.5 billion in the 5-year period to 2000 with growth rates ranging from 2% to 25% annually [13]. The value of the market created by enzyme technology is much higher, at around \$80–130 billion. In case studies by the Organization of Economic Cooperation and Development (OECD), the application of biotechnology has generally benefited by improved costs and sustainability, and

**Plant Molecular Pharming, Industrial Enzymes. Table 2** Summary of classes of enzymatic reactions and industrial uses

Enzyme class	Reaction catalyzed	Example of industrial enzyme/use
Oxidoreductases	Oxidation or reduction of substrate	Biocatalysis/fine chemical synthesis
Transferases	Transfer of a group from one molecule to another molecule	Transglutaminases/fine chemical synthesis
Hydrolases	Bond cleavage while water is added	Proteases, esterases/food, beverage and paper pulp
Lyases	Non-hydrolytic cleavage of bond and remove group from their substrate	Pectate lyases/food, beverage
Isomerases	Conversion of one isomer to another	Glucose isomerase/food, beverage
Ligases	Joining of two molecules at the expense of chemical energy	Synthetases/fine chemical synthesis

operational costs lowered by between 9% and 90% [12]. It is projected that the enzyme-based biotechnological industries will continue to grow, fuelled by trends including the demand for chiral chemicals, cost savings, emerging technologies, and sustainable industrial development [12, 17]. Novel enzymatic activities can also be generated without prior knowledge of detailed mechanisms [18], providing further impetus for enzyme-based industries.

Enzyme-based industries have certain advantages compared to traditional chemical manufacturing industries. For example, the production of many specialty chemical compounds, especially pharmaceuticals, relies on the use of chirality, or handedness, since many bioactive compounds and receptors show chirality. Chemical processes generally generate a mixture of compounds that are right-handed and left-handed, and separating them is problematic. Enzymes can be rationally designed to produce specific chiral molecules [19–21]. The worldwide market for single enantiomer drugs exceeds \$100 billion [22], making this search a worthwhile investment. In addition, enzymes being proteins are biodegradable and not harmful to the environment or difficult to dispose, unlike traditional chemicals which persist in the environment and are sometimes poisonous or bioaccumulate. For example, paper pulp was traditionally whitened using chlorine-based bleaches which are strongly oxidizing. New whiteners incorporate xylanases which are much less harmful than chlorine in the environment [23], and increase chlorine penetration, allowing less chlorine to

be used [24, 25]. Many chemicals used in traditional industrial processes are strongly acidic or basic which may skew the pH of the effluents or cause damage to containers. Some enzymes do require extreme pH or temperatures for activity (e.g., some proteases and enzymes from thermophiles) but most industrial enzymes have moderate pH and temperature requirements, overcoming these problems.

### Production of Industrial Enzymes

As the source of industrial enzymes varies, so do production procedures. Enzymes sourced from microbes, as well as transgenic enzymes produced in microbes, are usually grown in fermenters and can be extracted from the microbial cells. A preferred method is to secrete the enzyme into the culture media making extraction and purification much easier [26]. Microbial production utilizes microorganisms that have been modified and evaluated for safety and efficient production. Many have been used since historical times in the manufacture of fermented foods such as beer, cheese, soy sauce, and yogurt [13]. The first US patent granted for a transgenic microorganism with an industrial application was US Patent 4,259,444 following a Supreme Court ruling to Ananda Chakrabarty from General Electric Corporation for developing *Pseudomonas* strains harboring plasmids that could degrade aromatic hydrocarbons [27], and subsequently for mixed culture of *Arthrobacter* and *Pseudomonas* suitable “in the biological treatment of a contaminated

material including many persistent compounds of diverse chemical constitution” [28]. The ease of transformation of bacteria with plasmids and protection by patent law catalyzed the development of an industry of selected organisms producing proteins for various purposes.

Methods to optimize gene expression and enhance protein accumulation and purification continue to be developed. The industrial strains used to produce enzymes today are mostly proprietary and are selected for high production and accumulation levels. However, a disadvantage of bacteria is that proteins are not modified post-translationally and are often insoluble and accumulate into inclusion bodies. This can affect activity, especially when the protein is of eukaryotic origin, and methods to recover active proteins from this inactive conglomeration are often tedious [29].

Yeast is often the microbe of choice to express eukaryotic proteins, but hyperglycosylation of transgenic proteins has been observed. Fungal systems are relatively robust, but they have different metabolic pathways, post-translational processing, codon usage, and may form inclusion bodies [30, 31].

Following selection of an efficient microbial producer, the organism is grown in optimized conditions on media which may be solid or liquid. Most industrial enzymes are generally produced in 50–500 m<sup>3</sup> stirred fermenters. A major challenge with fermenters such as these is to maintain sterility, as contamination can cause loss of the entire batch. The integrity of the high producer must also be monitored to ensure that a lower-producing mutant does not outcompete the high-producing strain. But besides operational costs, capital costs can also be high, which makes this system quite expensive [32].

The same considerations apply to cultures of insect and mammalian cell cultures used to overcome the problems of eukaryotic protein expression seen in microorganisms such as glycosylation, folding, subunit association,  $\gamma$ -carboxylation, and cleavage [33]. For cell cultures, each production scale has to be optimized to the product and may vary from high-value, low-demand products being made in small multiple-unit reactors (flasks or roller bottles) and bulk products in large 10,000 L single-unit batch reactors to be cost-effective [34]. In fermenters, cultured mammalian cells are affected by shear forces and are also sensitive

to growth conditions such as pH and temperature, metabolites, and dissolved oxygen, which may affect product quantity and quality [5]. Such variation does not permit streamlining or standardization, and makes animal cultures more expensive to operate.

Enzymes such as catalase from liver and rennet from stomach can also be isolated directly from animal tissues, often as byproducts of the meat industry. Transgenic animal sources have the advantages of appropriate modifications of proteins [33], but problems of scale and costs of production, maintenance, and waste disposal. A serious concern with the use of animals is the risk of contamination of end products, especially when they may be used for human therapeutics or consumption, with animal pathogens such as viruses, mycoplasmas, and prions. More recently, Bovine Spongiform Encephalopathy (BSE) or mad cow disease was shown to be transmitted by contamination by small amounts of infective prions in blood and other animal tissues [35]. This had led to fears of contracting this disease by using protein products derived from animals.

Although plants were one of the first sources to be used to produce industrial enzymes such as papain from papaya and  $\beta$ -amylase from barley, only recently have they been developed as recombinant protein production systems [5, 36]. There are several theoretical reasons why plants may be one of the best sources for the long-term supply of exogenous enzymes including: (1) plants represent the least expensive method to produce proteins in general, (2) they do not require the large amount of capital for production compared to microbial fermenters, (3) production can be scaled up or down without major changes in infrastructure, (4) almost any plant in theory can function as a production system, (5) proteins can be targeted to specific compartments allowing for increased accumulation in the desired tissue with little interference in other tissues to reduce potential toxicity to the cell, (6) plants have convenient storage, transport, and processing of component materials; and (7) plants have the potential to combine lines with different enzymes through crossing [37–39].

There are other advantages. Many pharmaceuticals targeted for use in animals are toxic to animal cells. Plants do not share the same receptors and are capable of accumulating such proteins. Since plants do not

form inclusion bodies, the proteins stay soluble and can be purified more easily. Plants can further be engineered to produce protein in specific tissue, allowing the other parts of the plant to be processed to offset the cost of production. Further, the protein is thereby precluded from interfering with metabolism in other parts of the plant. Proteins can also be targeted to subcellular compartments, further reducing the risk of toxicity, as well as increasing production levels. This potential to express different proteins in different locations allows flexibility in storage and purification options [5, 40–42], and will be discussed in more detail below.

One obvious advantage is that plant-produced biologics are free of animal source tissue, thereby eliminating the fear of transmitting animal pathogens. The lower trophic levels plants occupy as producers is advantageous because it indicates that the energy input into plant growth is lower, and therefore the LCA impact is lower compared with microbes and cell lines grown in fermenters and animals on farms. Plants grown in fields require little more than air, rain, soil, and nutrients. Microbes and animals utilize plant material for growth, and unless they can utilize plant resources at 100% efficiency, they can never be as energy proficient as plants themselves. The low cost of production, ability to post-translationally modify proteins and clear growing, handling, and processing knowhow make plants valuable for industrial enzyme production.

In addition to the advantages of plants listed above, plant systems are particularly well suited to inexpensively yield large amounts of a desired product in a relatively small area. For instance, the cost of transgenic seed for extraction of  $\beta$ -glucuronidase in 1998 was estimated to be only \$0.20/kg [43], which was considerably less than bacterial cultures [44]. Moreover, because some plant tissues such as seeds can store proteins for years without loss of activity under ambient conditions, a ready supply of material can be manufactured into final form on an as-needed basis [45]. Propagation from stored seed, rapid scale-up, large volumes, and long-term storage are particularly advantageous for industrial enzymes. Low cost combined with the ability to use the raw material directly for industrial processes encourage development in this direction. These advantages have led to a recent

increase in use of this technology for the production of new biologics.

While it seems unlikely that one production system could meet all potential needs for the diversity of products, plants do offer some clear theoretical advantages over other systems. A summary of characteristics of different production systems is shown in Table 3.

### Production of Recombinant Proteins in Plants

There are a plethora of plants to choose from for heterologous protein production. The choice of the best plant type depends on how the characteristics of the final product complement the characteristics of the plant. Key factors include the ability of certain tissues to accumulate proteins, detrimental compounds such as toxins that may be produced in certain tissues that can co-purify with the protein products, the potential for the industrial crop to inadvertently mix with other food crops or weeds, the ease of purification of the protein from the plant tissue, and the potential to use the plant tissue directly eliminating the need to purify or extract the protein product. Table 4 lists some of the characteristics of different plant systems that can be used for protein production. While most plant systems can be used in theory, the associated cost can make this unsustainable for many industrial proteins. For high volume, the most cost-efficient system is with commodity grains. Grains provide the advantages of high protein content, feasibility for long-term storage, and the ease of downstream processing which give them great potential for future industrial protein production.

While plants are the least expensive source of biomass, they have not been developed to the extent of their microbial counterparts to accumulate proteins. The cost of producing the proteins is inversely related to the amount accumulated in the biomass so this has a direct bearing on the economics. In the past decades, most of the research on plants has focused on improving traditional uses of plants so there has not been much incentive to look at protein accumulation for use as a production vehicle for protein products. Currently high level of expression in plants is usually recognized at levels of 0.1% of the dry weight of the plant tissue. This leaves much room for improvement in the future.

**Plant Molecular Pharming, Industrial Enzymes. Table 3** Features of different industrial enzyme production systems

Production system	Bacterial	Fungal	Animal cell lines	Transgenic animals	Plant
Speed of creating transgenic plants	Rapid	Rapid	Rapid	Slow	Moderate
Capital cost to produce raw ingredients (fermenters, chambers)	High	High	Very high	Low to moderate	Low to moderate
Consumables (media and resources)	Moderate	Moderate	High	Moderate	Low
Processing cost	Moderate	Low	Moderate	Moderate	Low
Production issues	Contamination, maintenance of high producers	Contamination	Contamination, animal pathogens	Animal pathogens	

**Plant Molecular Pharming, Industrial Enzymes. Table 4** Characteristics of plant systems for the production of transgenic plants (Reprinted from [39]. With permission from Elsevier)

Crop	Advantages	Disadvantages
Wild species	Clearly distinguishable from crops	Low yield
		Outcross to native plants
		Little known about safety
Domesticated species	High yields	Potential to intermix with crops used for other purposes
	Infrastructure and experience exist	
Food	High margin of safety for human health products	Greater potential to intermix with food supply
Non-food	Less potential to intermix with food supply	Greater potential for toxic, antinutritional, or allergenic agents
Fresh tissue	Abundant biomass	Harvest/Transport/Storage
Seed or dry tissue	Harvest/Transport/Storage	
	High protein content	
Hydroponics, cell cultures	Limited exposure to environment	High cost
		Limited knowledge of product safety
Field grown	Low cost	Higher potential to intermix
	Infrastructure in place	
Modified food/feed grain designed for industrial applications	Clearly distinguished by color/shape	Not yet developed
	Non-transferable genetics	
	Low cost	
	Infrastructure and experience transferable from commodity crop	

Commodity plants currently used as a food, feed, or fiber source are being investigated as a production vehicle for industrial proteins. There has been public concern that use of food plants to produce industrial enzymes or pharmaceuticals may lead to inadvertent exposure to these products and cause safety concerns. Production of industrial or pharmaceutical compounds in organisms used in the food chain is far from new. In addition to the many native products isolated from animals, recombinant food organisms such as yeast or eggs play a major role in the production of pharmaceuticals and industrial proteins. There is also precedent in plants for species that produce both food and industrial products. Rapeseed is used primarily for the production of an industrial oil crop while canola seed which was derived from rapeseed, with subtle genetic differences, is used predominantly as a food crop. The key is to keep food and production streams separate [6] and failure to do so can create problems whether the organism is a traditional food or non-food source.

Current government regulations put plants on par with other production systems to prevent inadvertent products entering the food chain or harming the environment. While many of the industrial enzymes in use today are already in the food chain, these added precautions are necessary to limit exposure or can be used to protect against protein products that may not be in the food chain or have not undergone the rigorous or long-term testing needed to give confidence that there are no detrimental effects.

One concern often voiced by the public is that transgenic plants have the potential for dissemination of the transgene through pollen when grown in an open environment. The pollen may be ingested by non-target species, or hybridize with other plants. This situation has been recognized by regulatory agencies and there are strict controls on containment of transgenic plants and pollen. These include physical isolation and temporal delays as well as molecular containment strategies such as pollen and seed sterility and RNA interference have also been adopted to restrict dissemination [46–48]. These measures can lead to increased costs, but are necessary for safety and to allay public unease about transgenic crops.

The types of industrial proteins can include non-enzyme proteins and enzymes that have industrial use for food, feed, or pharmaceutical applications. This

also includes proteins used in the making of pharmaceuticals [49], including plantibodies [50–52] and edible vaccines [53, 54]. This may also include pharmaceuticals such as therapeutics and vaccines but these will not be discussed in this contribution.

Like pharmaceutical products, some industrial proteins may require appropriate post-translational modification and folding to be active. Most higher plants can accommodate this in a manner very similar to that which occurs with animal cells with minor modifications. Since plants do not form inclusion bodies, and since many proteins normally harmful to animal cells do not affect plant cells, plants are increasingly and successfully being used for their production.

One of the main limitations in using plants today to produce industrial proteins is the demand that cost must be extremely low compared to pharmaceuticals. This requires that the expression level be high. There are various options in terms of plant type, tissue, and intracellular location, allowing for great potential. However, this versatility also causes uncertainty in the early stages of developing a plant expression system. In addition, the plant's ability to accumulate a particular hydrolytic or oxidative enzyme has the potential for interference with the plant's metabolism and cause damage to the plant long before protein accumulates. While this has been seen in a number of cases, there have also been various ways to overcome this problem by tissue and subcellular targeting. In addition the use of thermophilic [55] and pro-enzymes [56] or the requirement for cofactors lacking in the plant have all been used to increase protein accumulation [57–59]. A list of enzymes from various sources (bacterial, fungal, animal, and synthetic) produced in plants is given in Table 5. This table provides a snapshot of the accumulation levels of specific proteins in selected tissues and the problems that investigators may have encountered by expressing the protein in the tissue.

### Options for Plant Transformation

The transformation technology used to express industrial enzymes can have a major impact on the accumulation of the recombinant protein. Therefore, it is important to select the type of transformation protocol that will best fit the application of the enzyme application. As the transformation process is discussed in

Plant Molecular Pharming, Industrial Enzymes. Table 5 List of industrial proteins produced in plants

Enzyme (gene)	Enzyme function; application in industry	Gene source	Host plant	Maximum expression level	Comments	Reference
$\alpha$ -Amylase	Starch degradation; food and beverages, biofuels, textiles, and paper industries	<i>Bacillus licheniformis</i>	Tobacco	0.3% total soluble protein (TSP) in leaf	Unaltered plant phenotype Secreted into intercellular space; extra complex sugar chains added; degradation products identical to native protein	[60]
$\alpha$ -Amylase		<i>Bacillus licheniformis</i>	<i>Nicotiana tabacum</i> SR1	0.4% TSP in seed	Constitutive expression; hydrolysis products identical to purified <i>B. licheniformis</i> $\alpha$ -amylase	[61] European Patent 044,9376
Thermostable $\alpha$ -Amylase		<i>Bacillus licheniformis</i>	<i>Vicia norbonnensis</i> L.	1 mU/mL seed supernatant	Seed specific USP promoter; accumulation in cotyledon protein bodies; post-translationally modified	[62]
$\alpha$ -Amylase OS103		<i>Oryza sativa</i> cDNA	<i>Nicotiana benthamina</i>	5% TSP in leaf	Viral infection causes mild chlorosis and stunting; moderate glycosylation of protein in plants	[63]
$\alpha$ -Amylase		<i>Bacillus licheniformis</i>	<i>Medicago sativa</i>	0.01% TSP in leaf	Unaltered phenotype	[64]
Bifunctional thermostable Amylopullulanase (APU)	Pullulan and amylose degradation; detergent industry	<i>Thermoanaerobacter ethanolicus</i> 39E (ATCC53033)	<i>Oryza sativa</i> L. cv Tainung 67	5.7% total soluble protein in seed	Unaltered phenotype; elevation in pullulanase correlates with decrease in amylose; amyloplast location; starch completely hydrolyzed upon heating	[65]
Aprotinin	Inhibitor of trypsin and proteases; medical and research uses	Optimized bovine aprotinin sequence	Maize	0.4% TSP in seed	Multiple copies in genome; ubiquitin promoter; protein accumulation in seed embryo; transgenic protein biochemically identical to native protein	[66]

Plant Molecular Pharming, Industrial Enzymes. Table 5 (Continued)

Enzyme (gene)	Enzyme function; application in industry	Gene source	Host plant	Maximum expression level	Comments	Reference
Aprotinin cDNA fusion with extension signal		Synthetic bovine	<i>Nicotiana benthamina</i>	7,100 trypsin inhibitory units/mg extract protein	Transient TMV virion transfection; product biochemically similar to native protein; large-scale production on 1.5 acres open field or 2,500 sq. ft. greenhouse yields 1 kg purified enzyme.	[67]
Arginine decarboxylase (adc) cDNA	Degradation of arginine; medical and research uses	<i>Datura stramonium</i>	Rice	2-fold increase in putrescine levels following stress removal	Under drought conditions, wild-type plants are severely affected, whereas transgenic plants have normal phenotype	[68]
Avidin	Irreversibly binds biotin; research uses	Chicken egg white	<i>Zea mays</i>	2.3% TSP in seed; 230 mg/kg seed	Partial to complete male sterility in high expressing plants; similar to native glycoprotein; stable during storage for over 3 months	[69]
Endochitinase (ech42) cDNA	Chitin degradation; research and agricultural uses	<i>Trichoderma atroviride</i>	Alfalfa	Up to 2,650-fold higher than control plants	Unaltered phenotype; high expression levels do not correlate with higher resistance to pathogen challenge	[70]
Endocellulase E1	Cellulose degradation; biofuels and paper industries	<i>Acidothermus cellulolyticus</i>	<i>Zea mays</i> L.	Higher levels in ER than mitochondria; max 2.0% total soluble protein	Targeted to ER and mitochondria; ER targeted E1-cellulase called "Spartan Corn 1"	[71]
Endo-1,4- $\beta$ -D-glucanase (E1 cellulase)	Cellulose degradation; biofuels and paper industries	<i>Acidothermus cellulolyticus</i>	<i>Zea mays</i> L.	6.1% (ER) and 5.6% (vacuole) TSP in seed	No apparent effect on growth; truncated catalytic domain accumulates in ER and vacuole; 16% TSP in single seed indicates high accumulation potential	[41]
Endoglucanase E1		<i>Acidothermus cellulolyticus</i>	Lemna minor 8627	0.24% TSP; 0.2 U/mg protein in fresh tissue	Unaltered phenotype	[72]



Endoglucanase E1	<i>Acidothermus cellulolyticus</i>	Tobacco	0.25% total soluble protein in leaf (apoplastic targeting)	Unaltered phenotype; stored seeds had 45% more activity after 1 year	[73]
Endoglucanase E1	<i>Acidothermus cellulolyticus</i>	Tobacco	Ammonia explosion (AFEX) treatment yielded 35% original activity	AFEX pretreatment is not a suitable method for releasing cellulase enzymes in transgenic plants	[74]
E1 endoglucanase	<i>Acidothermus cellulolyticus</i>	<i>Nicotiana tabacum</i>	1.35% TSP in leaf	Chloroplast targeting; normal growth and development; activity decreases with leaf age and upon dehydration	[75]
Cellobiohydrolase I (CBHI)	<i>Trichoderma reesei</i>	<i>Zea mays</i> L.	3.2% (cell wall) and 4.1% (ER) TSP in seed	Holoenzyme in cell wall; single seed levels of 17.9% indicate high accumulation potential	[41]
Exo-cellobiohydrolase I (CBHI)	<i>Trichoderma reesei</i>	Tobacco	0.11% TSP in leaf and 66.1 $\mu\text{mol/h/g}$ total leaf protein activity; 0.082% TSP in callus and 83.6 $\mu\text{mol h/g}$ total callus protein activity	Unaltered phenotype	[76]
Thermostable (1–3, 1–4) $\beta$ -glucanase (codon adapted)	<i>Bacillus</i> spp.	<i>Hordeum vulgare</i> L. (barley)	40 ng enzyme/ $2 \times 10^5$ protoplasts for codon modified constructs compared to none for unmodified constructs	Biolytic transformant; codon usage important for expression; unaltered phenotype; germination induced expression of enzyme in grain	[77]
$\beta$ (1–3, 1–4)-glucanase	<i>Bacillus</i> spp.	<i>Hordeum vulgare</i>	1.29 g/mg TS; 5.4% TSP in grain endosperm	Large variations in enzyme levels between transformants; levels stable for 3 years	[78]
Endo-1,4- $\beta$ -glucanase (EG1, cellulase)	<i>Trichoderma reesei</i> egl1	<i>Hordeum vulgare</i> L. Kymppi and Golden Promise	0.025% TSP in seed	Plant morphology normal but reduced seed setting in transgenic plants	[79]
Endoglucanase holoenzyme (E1) and catalytic domain (E1cd)	<i>Acidothermus cellulolyticus</i>	<i>N. tabacum</i>	1.6% TSP in leaf	Unaltered phenotype; apoplast targeting of catalytic domain achieves 500-fold greater expression than cytosolic full length E1	[80]

Plant Molecular Pharming, Industrial Enzymes. Table 5 (Continued)

Enzyme (gene)	Enzyme function; application in industry	Gene source	Host plant	Maximum expression level	Comments	Reference
1,4- $\beta$ -D-endoglucanase (E1)		<i>Acidothermus cellulolyticus</i>	<i>Solanum tuberosum</i> L.	2.6% TSP in leaf	Unaltered phenotype; dual crop applications: leaf targeting allows tubers to be used for culinary applications	[81]
Thermostable endo-1,4- $\beta$ -D-glucanase		<i>Acidothermus cellulolyticus</i>	<i>Nicotiana tabacum</i>	Not quantified	Targeting to chloroplast in vitro and in vivo	[82]
Modified endoglucanase cellulase (egl)		<i>Ruminococcus albus</i>	BY-2 tobacco suspension cells	0.1% TSP; 30-fold greater truncated form activity than endogenous cellulase	Unaltered phenotype; three forms (pre-form; mature form; truncated form); truncated form has highest expression	[83]
Hybrid (1,4)- $\beta$ -glucanase (cel-hyb1)		<i>Neocallimastix patriciarum</i>	Barley grain	1.5% total grain protein	Endosperm targeted; codon optimization leads to 527-fold increase in expression levels; stable during post-harvest storage	[84]
Catalytic domain endo 1,4- $\beta$ -D-glucanase (E1cd)		<i>Acidothermus cellulolyticus</i>	<i>Zea mays</i>	2.1% TSP in leaf and 0.845 nmol/ $\mu$ g/min activity; 2.08% TSP in root and 0.835 nmol/ $\mu$ g/min activity	Set seeds at maturity	[85]
Catalytic domain 1,4- $\beta$ -endoglucanase E1		<i>Acidothermus cellulolyticus</i>	<i>Zea mays</i>	1.13% TSP	Unaltered phenotype; apoplast targeted; successful conversion of corn stover into glucose following AFEX pre-treatment	[86]
Thermostable catalytic domain endo-1,4- $\beta$ -glucanase		<i>Acidothermus cellulolyticus</i>	<i>Oryza sativa</i> L. Japonica cv. Taipei 309	4.9% TSP	Unaltered phenotype; constitutive promoter; capable of hydrolyzing AFEX-treated stover	[87]
Truncated endoglucanase (t-egl)		<i>Ruminococcus albus</i>	Tobacco	0.5% TSP	Unaltered phenotype; cell disruption allows cell wall digestion to occur	[88]

Thermostable 1,4 $\beta$ -D-endoglucanase catalytic domain	Acidothermus cellulolyticus	<i>Arabidopsis thaliana</i>	26% TSP in leaf	No abnormal phenotype; apoplast targeting; activity and immunochemically similar to native enzyme	[89]
Thermostable 1,4- $\beta$ -D-Endoglucanase E2 and E3	<i>Thermomonospora fusca</i>	<i>M. sativa</i> L.	E2-0.1% TSP	Unaltered phenotype	[90]
		<i>N. tabacum</i> L.	E3 0.02% TSP		
		<i>S. tuberosum</i> L.			
Thermostable cellulases (Cel6A, Cel6B)	<i>Thermobifida fusca</i>	Tobacco	4% TSP	Homoplasmic, transplastomic plants using plastid-directed vector; not optimized	[91]
Recombinant hyperthermostable endoglucanase Cel5A	GenBank accession number At3g4890	Tobacco	5.2% TSP in leaf	Unaltered phenotype; chloroplast targeted; stable active enzymes	[92]
Chimeric chymosin (rennin)	Bovine	<i>Brassica napas</i>	0.5% (w/w) TSP	Seed targeted	[93] US Patent 7,390,936
Coumarate-3-hydroxylase (C3H)	<i>Medicago sativa</i>	<i>Medicago sativa</i> cv Regen SY	C3H levels 5% of wild type levels	No serious phenotypic impairment	[94]
Hyperthermo-philic $\alpha$ -glucosidase	<i>Sulfolobus solfataricus</i>	<i>Nicotiana tabacum</i>	0.04% TSP in leaf	Enzyme levels increase at maturity; unaltered phenotype	[95]
		cv. Xanthi			
Glycosyl hydrolase $\beta$ -glucosidase	<i>Sulfolobus solfataricus</i>	Tobacco	0.15% TSP in leaf	Enzyme levels increase at maturity; unaltered phenotype	[95]
ADP-glucose phosphorylase modified (Sh2r6hs)	<i>Zea mays</i>	<i>Triticum aestivum</i>	5-fold greater protein accumulation	Modified large subunits permit greater stability and yield; endosperm specific promoter	[96]

Plant Molecular Pharming, Industrial Enzymes. Table 5 (Continued)

Enzyme (gene)	Enzyme function; application in industry	Gene source	Host plant	Maximum expression level	Comments	Reference
ADP-glucose pyrophosphorylase modified large subunit (Shrunken 2 gene Sh2r6hs)		<i>Zea mays</i> L.	<i>Triticum aestivum</i> L.	91% more activity in the presence of 10 mM Pi	Transgenic wheat plants produced 38% more seed weight; 31% higher biomass; transgene stable after five generations	[97]
Human placental $\beta$ -glucosidase (GCase)	Degradation of glycosidic bonds; biofuels and research uses	Human placenta	Tobacco	750 U/kg seed	Seed viability totally impaired above 500 U/kg; taken up by human fibroblasts; free from immunogenic xylose and fucose	[98]
$\beta$ -glucuronidase (GUS)	Degradation of glucuronic acid residues; research uses	<i>E. coli</i>	<i>Zea mays</i>	0.7% TSP in seed	Functionally equivalent to native protein; dispersed in cytoplasm; single integration in transformant with highest expression	[99]
$\beta$ -glucuronidase (GUS) with $\alpha$ Amy8 regulatory and signal sequence		$\alpha$ Amy8 sequences from rice	Rice, tobacco, and potato suspension cells	40% total secreted proteins	Fusion to $\beta$ -glucuronidase (GUS); inducible by sugar; tunicamycin causes ER accumulation	[100]
Laccase I	Lignin degradation; biofuels, wood, and paper industry	<i>Trametes versicolor</i>	<i>Zea mays</i>	0.55% TSP	Variable expression levels; breeding and selection increased levels 20-fold in five generations; embryo-preferred promoter with cell wall targeting supports highest expression; germplasm background affects germination frequency	[110]
				4 ng/mg dry weight (T2) to 70 ng/mg dry weight (T6)		
Laccase		<i>Trametes versicolor</i>	<i>Zea mays</i>	0.20% of dried, defatted corn germ	Contains both water soluble and immobilized laccase; some laccase is inactive apoenzyme form	[58]

Laccase riceMaL and ricePcyL cDNA	<i>Melanocarpus albomyces</i>	Rice	13 ppm (riceMaL)	Endosperm targeted; seed production was normal; recombinant protein is biochemically similar to native proteins, but had lower kinetic parameters	[101]
			39 ppm (ricePcyL)		
Lipase	Recombinant dog gastric lipase	Tobacco	5% (vacuolar retention signal) and 7% (secretion signal) of acid extractable protein	Active glycosylated protein with similar properties to native protein; specific activity dependent on subcellular compartment; Normal leaf morphology	[102]
Lipase (rDGL)	Recombinant dog gastric lipase	Maize seed endosperm	Specific activities of 3 U/mg protein(grinding); 3.7 U/mg protein (defatting) and 9 U/mg protein (dry milling)	Comparison of relative efficiency of grinding, defatting, and dry milling of seed prior to selective extraction	[103]
Lipase	Dog gastric lipase	Tobacco	360 U/mg protein	Impact of subcellular targeting on glycosylation; transient expression system	[104]
Manganese-dependent lignin peroxidase (MnP)	<i>Phanerochaete chrysosporium</i>	<i>Medicago sativa</i> L.	0.5% TSP in leaf	Reduction in dry matter and height related to expression levels; yellow foliage; MnP expression segregates in sexual progeny	[64]
Manganese peroxidase (MnP)	<i>Phanerochaete chrysosporium</i>	Maize	15% TSP in seed	Cell wall targeting yields full length MnP; cytoplasmic targeting produces truncated products; seed-targeted promoter has higher expression levels and improved plant health outcomes over constitutive promoter	[59]
			3% TSP in leaf		
Anionic peroxidase cDNA	cDNA clone for the primary isoenzyme	<i>N. tabacum</i> ; <i>N. sylvestris</i>	>10× higher peroxidase activity compared to wild type	CMV35S promoter; chronic severe wilting through loss of turgor in leaves initiated at the time of flowering	[105]

Plant Molecular Pharming, Industrial Enzymes. Table 5 (Continued)

Enzyme (gene)	Enzyme function; application in industry	Gene source	Host plant	Maximum expression level	Comments	Reference
Phytase phyA2	Phytic acid breakdown; animal feed uses	<i>Aspergillus niger</i>	Maize	2,200 U/kg of seed	Embryo-specific globulin-1 promoter; different glycosylation pattern; stable over four generations; normal transgenic seed germination	[106]
Phytase		<i>Schwanniomyces occidentalis</i>	Rice	4.6–10.6 U/g fresh weight in leaves	Stable in silage for 12 weeks	[107]
Rationally designed phytase		<i>Aspergillus fumigatus</i>	<i>Triticum aestivum</i> L. (wheat)	4,777 FTU/kg seed flour	Vacuole accumulation despite apoplast targeting; unaltered phenotype	[108]
Secretory phytase (PHY)		Synthetic gene	Potato	40% more phosphate in transgenic plants	Trichoblast-specific promoter; healthy plants, but with altered leaf shape	[109]
Chimeric phytase ex::phyA		<i>Aspergillus niger</i>	<i>Nicotiana tabacum</i>	3.7-fold more phytase secretion and 52% higher P accumulation in transgenic plants	Presence of soil phytate essential	[110]
Phytase phyA		<i>Aspergillus niger</i>	<i>Triticum aestivum</i> L.	4-fold increase in plants with constructs with $\alpha$ -amylase signal peptide; 56% increase in plants with constructs without signal peptide; Phytase activity 3,000 FTU/kg	Endosperm, but not embryo accumulation; gene stability over three generations	[111]
Phytase gene		<i>Aspergillus fumigatus</i>	Rice Japonica var Taipei 309	130-fold increase in grain phytase level	Unchanged phenotype; coexpressed in endosperm with Phaseolus vulgaris ferritin gene and overexpressed endogenous cysteine-rich metallothionein	[112]

Phytase MtPHY1		<i>Medicago truncatula</i>	Arabidopsis	12.3- to 16.2-fold higher levels in root apoplast	Dry weight of transgenic plant upto 4.0 times higher than control and P content up to 5.5-fold higher; root-specific and constitutive promoters used	[113]
Phytase		<i>Aspergillus niger</i>	Tobacco	1% TSP in seed	Transgenic seed added directly to chicken feed shows improved nutritional quality	[114]
Phytase GmPhy		<i>Glycine max</i> L. Merr.	Soybean tissue culture	2-3 fold higher than controls	Novel phytase similar to purple acid phosphatases	[115]
Phytase cDNA		<i>Aspergillus niger</i>	<i>Nicotiana tabacum</i>	26% dry weight of leaves	Constitutive expression with secretion signal from tobacco pathogen-related protein S; differences in glycosylation compared with native protein	[116]
				14.4% TSP in leaf		
Phytase phyA		<i>Aspergillus niger</i>	<i>Brassica napus</i> (Canola)	600 U/g of multi-copy T1 seed; 103 U/g in single copy line	Unchanged morphology; seed-specific CruA promoter used; gene-dosage related expression; stable over three generations; no correlation between high expression and seed germination	[117]
Phytase phyA		<i>Aspergillus niger</i>	Soybean cell cultures	920 pKat M/g total soluble protein	Secretion and glycosylation may be necessary for activity; transgenic protein smaller than native protein, but has similar biochemical profile	[118]
Phytase		<i>Selenomonas ruminantium</i>	<i>Oryza sativa</i> L. cv. Tainung 67.	0.6 U/mg (appA); 1.4 U/mg (SrPF6) of TSP in seed; up to 60 times activity of control	No adverse effects; germination-inducible $\alpha$ Amy8 promoter and $\alpha$ Amy8 signal peptide; multiple copy number	[119]
SrPF6 ( <i>S. ruminatum</i> )		<i>Escherichia coli</i>				
appA ( <i>E. coli</i> )						
Phytase cDNA		<i>Aspergillus niger</i>	<i>Medicago sativa</i>	2.0% TSP	Un-glycosylated, but stable	[120]
Transglutaminase (rTGp)	Formation of peptide bonds; food industry	Rat prostate	<i>Oryza sativa</i>	0.15 U/mg h leaf	Ca <sup>2+</sup> -dependent enzyme	[121]

Plant Molecular Pharming, Industrial Enzymes. Table 5 (Continued)

Enzyme (gene)	Enzyme function; application in industry	Gene source	Host plant	Maximum expression level	Comments	Reference
Trypsin	Protein hydrolysis; medical and research uses	Bovine pancreas	Zea mays	0.025% seed dry weight	Equivalent to native enzyme. Levels suitable for commercial production; produced as zymogen	[56] US Patent 6,087, 558
Xylanase B xynB	Hemicellulose degradation; biofuels, wood, and paper industry	<i>Streptomyces olivaceoviridis</i>	Solanum tuberosum L.	5% TSP in leaf	Stable gene expression for several generations	[122]
Xylanase A xynA thermostable catalytic domain xynA1		<i>Clostridium thermocellum</i>	Rice	Not quantified	Normal plant phenotype; stable expression in seed and straw; activity in desiccated seed	[123]
Xylanase xynC-oleosin fusion		<i>Neocallimastix patriciarum</i>	Canola	2,000 U/kg seed (oil body of seed)	Fusion protein retains optimal temperature, Km, and specificity, but has reduced pH sensitivity	[124]
Xylanase Modified xynC		<i>Neocallimastix patriciarum</i>	Barley	0.004% dry weight of seed endosperm	Glub-1 promoter better than Hor2-4 promoter; protein stable during seed maturation, desiccation, and storage; 40% low fertility in one line;	[125]
Xylanase XYLII		<i>Trichoderma reesei</i>	<i>Arabidopsis</i>	1.2% (cytosol); 3.0% (chloroplast); 1.7% (peroxisome); 4.8% (chloroplast+peroxisome) total soluble protein	Unaltered phenotype; levels highest at flowering; dual targeting to chloroplasts and peroxisomes causes much higher levels than either compartment alone, although RNA levels are similar	[126]



Xylanase xynII					3.2% TSP in leaf	Chloroplast expression exhibit normal growth, but cytosolic accumulation affected transgenic plant growth	[127]
	Xylanase Thermostable, truncated xyn2	<i>Clostridium thermocellum</i>	Tobacco		4.1% TSP in leaf	Unaltered phenotype; proteinase II signal peptide used; enzyme enrichment following heat treatment	[128]
Xylanase Thermostable, truncated xyn2		<i>Clostridium thermocellum</i>	<i>Nicotiana tabacum</i> L. cv Wisconsin		Not quantitated	Clearance zone develops at 3 h	[42]
	xylanase (XYLD-A) and $\beta(1-3, 1-4)$ glucanase (XYLD-C)	<i>Ruminococcus flavefaciens</i>	Tobacco		170 $\mu\text{M}/\text{min}/\text{m}^2$ xylanase and 2,000 $\mu\text{M}/\text{min}/\text{m}^2$ glucanase in leaves	Unaltered phenotype; separate constructs; apoplast targeting; glucanase accumulated higher protein levels than xylanase	[129]
1-deoxy-D-xylulose-5-phosphate synthase DXS gene cDNA (CLA1)	Lignin structure modification; biofuels, wood, and paper industry	<i>Arabidopsis thaliana</i>	<i>Arabidopsis thaliana</i>		Higher or lower levels than wild type	Normal morphology; variable levels of isoprenoid biosynthetic pathway products correlated with increased or decreased enzyme level	[130]

detail elsewhere, only a brief synopsis of how it impacts industrial enzymes is summarized here.

Both stably transformed and transiently expressing plants have been used to express proteins [131]. In stable transformation, the foreign DNA can be targeted to the cytoplasm or a number of different intracellular locations such as the nucleus [132] or into plastid genomes, usually the chloroplast [133]. Mitochondrial targeting is not as well established and has not been pursued in this context. Organelle transformation provides the advantages of high copy number and the transgenes are not passed on by the pollen [6]. This method however is not yet applicable to many types of plants. Stably transformed plants are time consuming to characterize and generate, but once produced, they can be stored as seed. This allows for a ready source upon demand. Transiently expressing foreign DNA can be inserted into somatic tissue with the purpose of short-term expression using viral vectors delivered using *Agrobacterium* or biolistics [134–136]. Transient expression is useful when a protein needs to be expressed at short notice.

### Selection of Plant Species for Transformation

One consideration for industrial protein production is the type of plant used as the production vehicle. The options include: plants grown for their vegetative tissue for large volumes of biomass; plants harvested for grain for their enriched protein and facile storage characteristics; well-established cultivated crops where much is known about growing and processing; wild species that have little use today, making them distinct; food crops that have Generally Regarded as Safe (GRAS) status and pose no safety threat to the crop itself or host protein but have the potential for intermixing with food crops; or a non-food crop with decreased concerns about intermixing with food crops but with greater potential to have compounds that are detrimental or untested with regard to human safety (see Table 3).

Food crops are known to be safe when consumed and they have well-established procedures for growth, harvest, and storage. In cases where the final product may include some or all of the plant tissue as well as the recombinant protein, this has a significant advantage and direct applicability in the case of many industrial enzymes that are used in the food and feed industry.

For example, maize (*Zea mays*; corn) is well accepted as a safe product (GRAS), and is widely used in food, feed, and industrial applications today [137]. The production cost of maize is very low, and the infrastructure can handle large or small acreages for industrial or pharmaceutical products. Storage and transport of seed, and protein purification from flour are compatible and flexible with current practices without special handling. There are no known agents in maize that generally interfere with protein purification. Finally, the grain can be processed with little or no heat inactivation steps without affecting the protein's properties [4, 138, 139].

In addition, maize has an added advantage in that the kernels can be mechanically separated to yield a germ fraction with enriched protein and an endosperm fraction with enriched carbohydrate [137]. This facilitates use of the carbohydrate fraction for industrial applications such as ethanol production [140] or feed. In this way not only is the cost of the raw material reduced but the waste products are conveniently handled as well.

A disadvantage of maize is the fear of inadvertent mixing with the food supply. Intermixing potential can be handled by management practices but will need to gain the public's confidence. While there may be claims that no food source can ever be used to produce pharmaceutical products, it is common knowledge that both eggs and yeast are used to make pharmaceuticals. Not only is there no public outcry of intermixing in these instances, but their perception is that these production systems are distinct from the systems used to make food. This is the perception that food crops must earn as well. Maize as well as other plant systems must build an infrastructure dedicated to industrial protein production, which is as distinct from food production as edible eggs are from vaccine production. Furthermore, this must also be perceived by the public to reduce fears.

One advantage of non-food crops for industrial protein production is that they are less likely to be mistaken for food crops and therefore unlikely to be inadvertently mixed with the food supply. The disadvantage is that non-food crops must be assessed for toxins, allergens, or anti-nutritional agents that may accompany the recombinant protein. For industrial proteins where little purification is performed to keep

the cost down this can be a considerable problem if the protein precursors are to be used in food or feed applications.

Another question that is relevant to selecting a crop for recombinant industrial protein production is whether a cultivated species or wild species is preferred. Cultivated species show “domestication syndrome” and have several advantages for humans than that their wild relatives exhibit through husbandry and selection over thousands of years [141]. Although non-cultivated species have not undergone selection for higher yields or been subject to agronomic practices in past centuries, there may be yield advantages in crossing domesticated plants with wild relatives, with increase in yields of up to 50% in tomato fruit [142]. Finally, the impact of relatively unknown wild species on product safety is unknown. Determination of the degree of this impact would undoubtedly require extensive effort and time, which may be limiting factors for industry.

A further choice is whether to use an open-pollinated or a self-pollinating plant. Self-pollinating plants have the advantage of a lowered risk that pollen will unintentionally transfer onto other plants of the same species. Controlled pollen shed of open-pollinated crops can be used to help alleviate this concern by either physical or genetic means to prevent out-crossing onto weedy species or related food or feed crops [46–48]. The likelihood of gene transfer from engineered to wild plants depends on sexual compatibility, flowering time, and pollen transfer distance between the engineered and wild varieties [143]. This choice is made based on the growing location and the product required.

With a variety of species to select from and the wide variety of products that are possible, it is highly unlikely that any one system will work best for each of these steps and therefore, it is important to select the plant system that best suits the product. Since it is impractical to have thousands of different production systems, it is preferable to adapt a given system to the needs of various products.

Fortunately, some common features exist that will apply to most products enabling a few systems to accommodate most products. The key features include a potential for low cost of goods, maintenance of protein integrity, flexibility with regard to time and

temperature for harvest, and maintenance of product safety and environmental safety [144, 145]. Production systems are discussed below as to how they relate to the overall efficiency of the system as well as to the regulatory aspects.

### Selection of Plant Growth System

Transgenic plants can be grown in the traditional manner in an open field or in a contained chamber such as a greenhouse [146]. Row-grown transgenic crops are subject to USDA regulations on buffer zones to limit pollen spread and minimize the risk of the intermixing of food crops with “pharma” crops. If the protein source is green tissue, the plants may be harvested before flowering to limit pollen spread. This is also the case when transient transfection is used to express proteins in leaf and other green tissue. When row plants are required to flower and set seed for protein production, such as is the case for grain-targeted proteins, adequate precautions must be taken to avoid pollen transfer. Row plants are capable of higher product accumulation since they are capable of growth to higher biomass [38].

There are strict regulations and permit requirements imposed by APHIS, the branch of USDA with responsibility for animal- and plant-related services ([www.aphis.usda.gov](http://www.aphis.usda.gov)). A practical solution proposed for field growth is to dedicate areas to growth of transgenic plants, or to grow transgenic plants in locations within processing areas, such as cellulose-producing plants in a bioethanol production area [140]. This way the material is grown where it is utilized, cutting transportation costs. Alternatively, the transgenic plant can be isolated in greenhouses in soil or in liquid media. This solution is more expensive, but has the advantage of physical isolation of the transgenic plant from the environment.

The potential for inadvertent transfer of the transgene to non-target plants is considerably reduced by containment in a chamber, but production is necessarily limited by cost considerations. Within a chamber, plants can be grown in soil or in liquid growth media where exudates containing the protein of interest may be continuously produced and has similar advantages to plant cell cultures where proteins are secreted into the culture medium [147]. Scalability may also be an

issue. Exudates from roots also facilitate production when secreted into fluid growth medium [40], but chambers for contained growth in liquid medium can be capital intensive and require sterility and media-related expenses. On the other hand, secretion of transgenic proteins into guttation fluid provides a convenient, but labor-intensive method of recovery of secreted proteins. Proof-of-concept was shown with three transgenic proteins in tobacco, [42]. Using this “phyllosecretion” technology, Komaryntsky and colleagues engineered three proteins from different genetic backgrounds (bacterial xylanase, jellyfish green fluorescent protein and human alkaline phosphatase) into *N. tabacum*. The proteins were fused to endoplasmic reticulum signal peptides targeted to leaf apoplast and released into guttation fluid in plants maintained in high humidity conditions. Guttation fluid has lower overall protein content than apoplast fluid, and is also released throughout the plant’s life. In addition to continuous production, the process is non-destructive of plants. Levels of up to 1.1 µg/g of leaf dry weight per day, comprising up to 3% TSP, were recovered from the guttation fluid using simplified downstream processing. Although tobacco is not an ideal plant for guttation fluid production, other plants such as tomato and some grasses are highly susceptible to the production of large quantities of guttation fluid and may provide alternative targets for phyllosecretion.

### Optimization of Heterologous Protein Production

Because industrial enzymes require a very low cost of production the most critical factor determining the system of choice is the level of accumulation that occurs in the selected tissue. The optimization of expression of heterologous proteins in plants has been studied for various purposes: to improve nutritional value, insect resistance, salt and drought tolerance, and product quality. These other applications however are not nearly as demanding for the level of expression required as for industrial enzymes. Key steps to increase heterologous protein expression are the choice of plant, tissue location, various manipulations of the promoter, codon usage, and compartmentalization of the protein.

After selecting the type of plant, a determination has to be made about which target tissue is best to express the protein. Location of expression is guided

by many factors such as the nature of the protein, whether the protein is to be used directly or purified, and accumulation levels desired. Often, strong constitutive promoters such as the CMV 35S promoter are used, and promoter analysis by site-specific mutations has allowed delineation of sequences that modulate expression in tissue-specific locations of this promoter [148]. *Cis*-acting elements in a green-tissue-specific rice promoter  $P_{D540}$  acting as activators or suppressors of activity were identified, thereby facilitating control of protein expression in different tissues [149]. Manipulations such as the use of the embryo-specific maize globulin-1 promoter also allow accumulation of proteins in specific locations within tissues [150]. Expression of the heterologous protein can also be controlled by the use of inducible promoters, and a search of patents reveals a plethora of such inducible promoters. A list of promoters in cereals is compared in [151]. Other promoter permutations including the use of inducible promoters, stacking transcriptional units, synthetic bi-directional promoters, global regulatory sequences to recruit transcription factors, and other such approaches have been studied to enhance transcription and are reviewed elsewhere [152].

Further, heterologous gene sequences should be optimized for the plant type and location. For example, chloroplast codon usage is similar to that of prokaryotes, whereas nuclear codon usage varies from plant to plant. Also, RNA silencing is a feature of many plants, and some viruses produce suppressors of silencing, and heterologous protein sequences are often not expressed at high levels because of RNA silencing [49]. This silencing can be turned off by the expression of the heterologous gene together with a suppressor of silencing [153, 154].

In addition to the choice of tissue location, highest expression levels can be obtained when the protein is directed to specific subcellular compartments and especially so if the protein is an enzyme that would interfere with normal cellular activities [38, 41, 152, 155]. Subcellular targeting is critical for accumulation and protein integrity of hydrolytic enzymes. Cell wall targeting allowed expression of full-length manganese peroxidases, but cytoplasmic targeting resulted in truncation of the peptide [41, 59]. In addition, a seed-preferred promoter allowed high accumulation without negative effects on plant health [59]. Interestingly, targeting xylanase to two subcellular locations – the

chloroplast and the peroxisome – accumulated 240% more enzyme than the chloroplast and 160% more than the peroxisome alone. Highest levels also accumulated during flowering time [126]. These are empirical observations, and such permutations may benefit studies of high-level expression of heterologous proteins. Medrano et al. have developed a transient expression system using *Nicotiana* to assess construct efficiency [156], which may be helpful.

Finally, decisions on the expression of a heterologous protein as a fusion or as a free protein depend upon target location and levels of expression desired. For example, fusions to oleosins accumulate in oil bodies in *Brassica napus* seeds [157], and fusions to proteins previously shown to stably accumulate in plant cells may have the effect of stabilizing heterologous proteins as well [158]. Fusions to the C-terminus of ubiquitin also have a stabilizing effect resulting in 10-fold increased levels [159]. Proteins accumulated at high levels may be subjected to cellular protease activity, thus using cell lines with lower protease levels may help stabilize the heterologous protein. The expression of some proteins, such as hydrolytic enzymes, can be detrimental to the cell. The strategy for the protease trypsin was to express it as a zymogen, allowing sufficiently high levels for commercial production, marketed as TrypZean [56].

### Product Recovery from Transgenic Plants

The single most important consideration for recovery is whether the protein produced in plants can be used in crude form in the plant extract, or has to be purified prior to use. Obviously, crude extracts are considerably less demanding to make, but the end use of the protein determines the level of purity required. Purification of transgenic proteins is an expensive prospect, regardless of the system, accounting for about 94% of the production cost in the case of maize [44]. Decisions on recovery should form an integral part of assessing production options based on levels of expression, costs, and sustainability. These decisions are based on the nature of protein to be expressed, and include transient or stable expression, type of plant, targeting, and modifications required.

A variety of elegant solutions have been used to overcome the problems of purification, but vary in

costs and efficiency. For example, roots have been used to secrete proteins into the growth medium, and guttation liquid produced by plants also provide a useful medium for secretion-based isolation of the transgenic protein, as the growth or secreted liquid can be recovered and filtered for protein recovery [42, 49]. However, they are considerably more labor and energy intensive due to growth in liquid medium as discussed above, and recovery varies based on loss due to dilution factor and protein stability in an extracellular environment [160].

When whole plants are used as production systems, the recovery process depends on the tissue being used for expression. Transient expression in roots or leaves generally requires processing of fresh wet tissue. If stable expression is used, the material may be wet or relatively dry depending on the tissue used. Proteins extracted in fresh tissue are generally unstable and have to be recovered immediately, whereas seed-expressed proteins have shelf-life as long as 3 years [45]. Grain can be subjected to either wet or dry milling followed by separation steps. These include fractionation to obtain enrichment of germ or endosperm in the case of seed, or subcellular compartments targeted for protein accumulation, protein precipitation, adsorption, chromatography, and diafiltration [44]. Methods have been developed to simplify recovery such as post-harvest protein expression using stress-inducible promoters [45, 161] and oleosin-partitioned proteins that can be recovered easily following oil-water separations [124, 162, 163]. A detailed comparison of the economics, processing, and regulatory constraints associated with the most common plant production systems is provided in [145].

In the ideal case, commodity plants that are used for industrial applications would also express the transgenic protein. These transgenic proteins can then be used directly in the industrial process without purification assuming the protein can survive the processing steps. In this best-case scenario, no other inputs are needed to produce, purify, or process the protein leading to lower cost and less detrimental impact on the environment.

**Case Study: Economics of Cellulases Produced from Transgenic Plants** With the current state of technology for biomass conversion, the overwhelming enzyme

requirement is for cellulases: endo-cellulase, exo-cellulase, and glucosidase [164]. The specific activity of most cellulases is quite low [155, 165] and much effort has focused on increasing their activity levels. However, even with improved enzymes and improved methods of production, the amount of cellulase required to deconstruct the volumes of biomass necessary for 30% replacement of gasoline are in the millions of tons. It has been estimated that 36 billion gallons can require as much as 3.6 million metric tons of cellulase per year [166]. This is an unprecedented challenge in terms of the amount of enzymes and the extreme low cost that is required.

The bioprocessing challenge for ethanol is how to deliver these extremely large volumes to a saccharification facility at an unprecedented low cost. This represents potentially the single greatest application of industrial enzymes.

Currently, cellulases are produced by fungal and bacterial systems, and are a costly component of ethanol production [167]. Plants offer the potential for a production system that can meet the low cost and high volumes required. To do this however, the level of expression needs to be extremely high and the choice of tissue needs careful consideration. There have been many attempts to express cellulase in many types of plants and these have been reviewed elsewhere [37]. Protein and tissue stability, tissue fractionation, protein extraction and formulation, storage, as well as transportation add to this cost and must be considered.

To achieve targeted production cost targets for biomass enzymes the following considerations must be employed: (1) eliminate transportation cost by integrating enzyme processing into biomass conversion facility; (2) minimize fractionation/extraction cost of transgenic material; and (3) reduce the contribution of transgenic plant material to enzyme production cost by capturing plant biomass value through byproduct credits or cellulose. It has been suggested [166] that in order to keep the enzyme cost down, plant production systems that accumulate cellulase in the normally unused or low-value portion of the plants can be competitive when the other parts of the plant are harvested for their traditional use. The obvious example is when the leaves of crops are used to produce the cellulase and the grain is used for food, feed, or other industrial applications. Using the grain for industrial applications

is less of a regulatory burden than if the grain is to be used for food or feed. However, according to the US Food and Drug Administration Statement of Policy (<http://www.fda.gov/Food/GuidanceComplianceRegulatoryInformation/GuidanceDocuments/Biotechnology/ucm096095.htm>) this is still an option, "If plants (or materials derived from plants) used to make nonfood chemicals are also intended to be used for food, producers should consult with FDA to determine whether the non-food chemical would be a food additive requiring an authorizing regulation prior to marketing for food use." These guidelines may change and may vary between countries; therefore, current regulatory practices must be consulted and followed in each case.

A case has been proposed for a fully integrated and synergistic system of ethanol production using maize. Maize grain today is the major source of ethanol production in the USA [168]. Using the stover to make cellulosic ethanol has been proposed as a convenient, economical, and sustainable way to make cellulosic ethanol alongside grain ethanol allowing for lower transportation costs and synergy in the ethanol facilities [170].

Taking this approach one step further, it has been suggested that the leaves themselves can be used to generate the required enzymes [155]. This has great potential if the enzymes can survive in the processing steps and can reach the target levels.

Another option has been proposed using the germ fraction of the grain [140]. In this case the enzymes could be made in the germ which is separated in the dry milling prior to using the endosperm for grain ethanol. The acreage required to grow crops to produce this amount of enzymes has been modeled previously considering the proximity limitations of the lignocellulose biomass to the ethanol facility to avoid large transportation cost. This study demonstrated that if cellulase levels were 0.1% of the dry weight of the seed (1% of germ dry weight), this was more than sufficient to keep the acreage of the cellulase crop less than the acreage needed to supply the lignocellulose biomass. Additional models suggest that when expression levels reach 4% of the dry weight of tissue (0.4% weight of the grain) it may be possible to add the germ tissue without any fractionation [166]. Direct delivery of plant tissue can be the system of choice by eliminating

extraction and purification costs. In addition to the cost of production, this approach allows the entire plant to be better utilized without additional acreage or input for growing the plants. There is also no additional stress on the environment due to making or processing the enzymes and this has the potential to meet the current volumes of cellulase projected at a cost below current targets.

The main reason why this is not in use today is that cellulase has not enjoyed the levels of expression required for the models above. However, there is every reason to believe that expression levels will continue to improve in plants as it is a relatively young science compared to microbial production. In maize specifically it has been reported that levels as high as 16% of the total soluble protein were observed [41] and more recently levels of >1% of dry weight in the germ fraction has been achieved [169].

### Regulation of Growth and Use of Transgenic Plants

Assuming that one can create plants having industrial enzymes with the characteristics and cost benefits that are desirable, there is still the need to grow these on a commercial scale. The process of growing transgenic plants is highly regulated including those grown for research purposes. However, for research purposes, cost is not a primary concern and the environmental impact is usually minimal due to the small acreage. However, this changes dramatically during scale-up for commercialization. Assuming a yield of 1–10 kg industrial enzyme per acre would require 100–1,000 acres for a relatively low volume of industrial enzymes but this can increase 10- to 1,000-fold for larger volume enzymes. Therefore, the potential for certain individual enzymes to be greater than 100,000 acres creates regulatory scenarios that are much more complex than those addressed in a research environment.

Regulatory systems are a social issue and vary in different countries but they all address human and animal safety as well as environmental implications. They must also address perhaps the most controversial issue, public acceptance. The technical aspects of regulatory concerns are discussed below with the understanding that every country will have its own interpretation and standards for what is acceptable.

### Product Safety

The first concern for any product is the inherent safety of the active ingredient. Having a production system different from the native host does not usually change the inherent properties of the enzyme itself. Factors such as exposure to humans, dosage, toxicity, allergenicity, and whether or not the proteins are already a part of the food chain are considered. It is not the intent of this entry to review the regulatory process in detail but rather to point out the difference when using plants as the production vehicle as opposed to other sources. Therefore, the focus is not on the inherent safety of the enzyme but what safety factors arise when produced in plants and what additional challenges are presented with plant production.

For products that have already undergone regulatory approval, it is critical to show, utilizing empirical data, that there is equivalency with the plant-produced process. In many cases plants can produce functionally and chemically equivalent proteins to those made in the native hosts. However, there may be exceptions which in turn can lead to difference in protein structure or function. These differences may be subtle such as a small signal peptide intended to target the protein to the vacuole that is not cleaved and instead retained in the final protein sequence, or carbohydrate structures could be added where none existed before. These types of changes may have no impact on function and may be acceptable in commercial products.

In contrast to that above, some changes may result in proteins that have altered functions or altered safety profiles. As an example, many industrial enzymes are bacterial in origin and therefore are not normally glycosylated proteins. If these bacterial proteins happen to have a glycosylation site that a eukaryote can recognize then that can create a challenge. The possibility exists that the enzyme will be glycosylated when expressed in eukaryotes and potentially lead to a change in function. Fortunately there is little evidence to suggest that glycosylation itself will change enzyme function unless the carbohydrate is added to a key amino acid either in the binding site or catalytic site. An example of this is the bacterial enzyme  $\beta$ -glucuronidase (GUS). GUS protein loses activity due to secretion-specific N-glycosylation of a key amino acid in the protein [171]. Contrary to this

observation, other proteins can be glycosylated such as the bacterial protein organophosphate hydrolase with no significant effect on its biochemical activity [57].

Glycosylation patterns in plants can also be different from those observed in other eukaryotes. Plant glycoproteins contain the same mannose backbone structure found in animal glycoproteins but the additional sugars added to the backbone are usually less complex in plants than animals. Another difference is in  $\beta$ -1-4 linkage versus  $\beta$ -1-6 linkage and the appearance of xylose. In one case the animal protein trypsin was produced in plants and showed evidence of O-glycosylation whereas it is not detected from the native porcine or bovine source [56]. Despite these changes no functional difference in catalytic activity has been observed.

In addition to catalytic activity, there is also the concern that the enzyme may be altered in a way that affects its activity in the desired application. There is little evidence of this for industrial enzymes but there are examples for where carbohydrate structures on proteins can affect their pharmacological properties. In the case of certain pharmaceutical proteins, sialic acid is a terminal sugar on the glycoprotein and leads to a longer half-life in the blood [172]. Since plants do not add sialic acid this leads to an alteration in clearance time in the blood [173]. On the other hand, the plant carbohydrate sequence for antibodies is also critical *in vivo* but the altered plant sequence appears to work as well as the animal carbohydrate sequence [174]. This demonstrates the need to test the industrial enzymes in the desired application when the composition of matter is different than the native source.

Demonstrating functional equivalency is still not enough from a regulatory standpoint. The addition of carbohydrates has been implicated in a number of studies for allergenicity. This raises the theoretical question of whether plant glycosylation can lead to allergenicity. On the surface this seems highly unlikely since plant proteins are eaten routinely, and plant glycoproteins would seem inherently safe. Therefore, just adding a plant carbohydrate does not make the protein allergenic. On the other hand, there have been reports showing that the carbohydrate sequence of pollen glycoproteins [175] is responsible for an allergenic reaction. Therefore, while generalizing that plant

carbohydrates are allergenic is misleading, it is important not only to check whether there is glycosylation but also to find out how this may differ from the native source. While there are no current examples where the addition of a plant-specific glycosylation to a transgenic protein has led to an increase in allergenicity there is still a theoretical concern.

### Safety of Host Proteins

While inherent activity is a functionality concern, the addition of extraneous host material is also a regulatory issue. Since most industrial enzymes cannot be purified because of cost restraints, the host material must also be shown to be safe in the final product. For this reason production in plants that already have a proven history of safe use is a great advantage. Certain crops that are known to produce toxins, allergens, antinutritionals, or carcinogens present additional difficulties as hosts. Crops already known to have GRAS (generally regarded as safe) status will be at an advantage because of inherent safety of the crop. In the best case a GRAS enzyme can be produced in a GRAS host greatly reducing the regulatory burden [176].

### Environmental Safety

The environmental impact of protein production from a regulatory standpoint can be a concern based on the additional acreage. The consequence of additional inputs, displacement of food crops, and the lack of containment leading to inadvertent exposure to animals and humans must be addressed [177].

Dedicated cropland for the sole purpose of producing industrial enzymes will in most cases be insignificant compared to the acreage already under development for current uses, if produced in tandem with commodity crops. Since expression levels must be high to keep costs low for industrial products, this translates into only a very small percentage ( $\sim 1\%$ ) of the acreage for even large-volume products. The exception to this scenario is if the by products can be utilized for other, value-added purposes, such as for biomass or feed. In such cases, there is no additional impact either in acreage or inputs.

Of greater concern is the issue of containment and inadvertent exposure. Regulations to evaluate containment of transgenic plants are similar in concept to



those for other hosts, but specifics are very different. One of the areas historically concerning the production process of industrial enzymes is the occurrence of allergic reactions developed by workers in the production facility. This has led the industry to safer forms of controlling small-particle exposure during microbial production. The concern for plant is that pollen from production fields can act as a carrier of the protein and lead to exposure similar to that observed for microbial production. For this reason, it is suggested that expression of the protein be specific to tissues other than pollen, thereby eliminating this concern.

Field production has the potential to affect wildlife where the crop is grown. Unlike many applied pesticides, transgenic plants are generally non-toxic to wildlife. This may reduce the environmental threat as opposed to many chemicals. The specific protein produced still may have some unwanted effects on wildlife and this aspect needs to be addressed, particularly for endangered species that may be present in the production field. Products that require larger acreage will create more concerns.

The final evaluation and perhaps the most controversial is the potential impact for the industrial enzyme for inadvertent exposure into the food chain. This may occur for several theoretical reasons, including mislabeling of seed, spills during transport, or pollen dissemination into food crops or wild relatives. Pollen dispersion has received the most attention as the other possibilities of inadvertent exposure most likely because the other potential sources are similar to that of other host production systems. To control the flow of pollen several different methods can be employed. These range from male sterility systems and physical isolation of the crops from compatible agricultural crops or weeds. In addition, systems have been proposed for growing transgenic crops within an industrial crop zone thereby further reducing the possibility of inadvertent mixing [140].

The underlying assumption for the production of transgenic proteins produced in plants is that they must be kept at all times isolated from food crops. This is essential for proteins that have the potential for detrimental effects on the population. However, many industrial proteins are used in food processing or derived from material already in the food chain. In these cases the transgenic proteins have the potential to

be deregulated. After demonstrating that there is no danger to human safety the strictest of containment conditions may be dropped although some containment may still be desirable, either from a regulatory standpoint or from a commercial necessity.

### **Plants as Sustainable Sources of Industrial Enzymes**

Plants have the potential to provide a sustainable source of industrial enzymes. Most plants can be transformed stably or used to express transiently heterologous proteins at levels that can be used for industrial production. The key sustainable feature of plants is that they are a renewable resource. They do not require intensive efforts for growth and maintenance of sterility, although some may require containment. Even so, the infrastructure and capital investment for growing plants is considerably less than that for fermentation of microbes. In addition, plant waste can be disposed of without intensive treatment, unlike effluents from fermenters. Energy resources for plant growth are also lower than those for maintenance of temperature and sterility of fermenters. In addition, formulations of large volumes of media for culture represent a large input of water, often a limiting resource. Cooling of large fermentation chambers is also energy intensive. These considerations are less severe with plants. Although plants do need regular and sustained watering, they can be grown in traditionally non-irrigated areas or in irrigated land where the input can be spread out using efficient drip hoses and watering procedures over a long period of time, alleviating the need for vast quantities of sterile water at short notice. The water does not need to be sterile, which reduces the energy burden and thereby the LCA.

The projection of 2–25% annual growth of enzyme requirement indicates a massive increase in enzyme production if supply is to keep pace with demand. Large volumes of enzymes imply that a number of fermenters have to be constructed to produce microbial- or cell-culture-based enzymes, or a shift of paradigms to a more efficient supply. Plants can provide the large volumes of enzymes needed with relatively low capital investment, and may represent the only really low-cost option for providing the immense requirements of industrial enzymes anticipated with projected

growth. Depending on the level of purification required, the enzymes can either be used without purification (such as for the production of biofuels) or undergo processing (such as for pharmaceuticals). This decreased processing obviously lowers environmental impact of the procedure. Tissue from plants that are not target for protein accumulation can be processed or disposed of such as to provide a financial buffer to the industrial enzyme production aspect.

The key to using plants for the production of industrial proteins is the increasing expression of transgenic proteins to levels commensurate with economic recovery of protein. A search of the Sigma Aldrich Chemical Company for transgenic-plant-derived products showed that proteins expressed in rice (avidin, lactoferrin, lysozyme), corn (avidin, trypsin), and tobacco (tissue factor proteinase inhibitor II, bovine aprotinin) are commercially marketed. Cell Sciences ([www.cellsciences.com](http://www.cellsciences.com)) produces over 25 cytokines and growth factors from barley endosperm, especially marketable as they are animal, bacterial, and viral-free. These are fine chemicals produced with high purity, but evidently are also commercially viable, providing a proof of potential. Increasing accumulation levels in plant tissue is an important issue for commercial success.

Storage of enzymes is also energy consuming when produced from microbial fermentation. The protein is usually lyophilized for storage; an energy-intensive process requiring freezing and desiccation simultaneously. Green tissues from plants have to be processed immediately, frozen or dried for protein stability. However, proteins expressed in seeds have been shown to maintain stability, even at room temperature, for at least 3 years without noticeable degradation [45]. This decreases the LCA and increases sustainability, as well as facilitating rapid response to spikes of increased demand.

Plant-based production also results in less waste. The unused portions of the plant body can be funneled into other uses, such as ethanol from biomass, and there is little waste from the recovery process compared to effluents from fermenters, lower waste disposal requirements, and lower net production of greenhouse gases, which is better for environmental sustainability and society.

## Future Directions

Plants have historically contributed to industrial processes from dyes and tannins for fabric and leather to drugs for healthcare and pharmaceuticals. The first generation of plant-derived recombinant proteins is now commercially available and the prospects of more products is in the pipeline with many groups working on expressing high levels of laccases, cellulases, plantibodies, and pharmaceutically important proteins. In some cases, enzymes can be directly delivered in the plants, such as cellulolytic enzymes expressed in plants to improve their degradation for production of bioethanol. Alternatively, enzymes can be expressed at high levels and isolated for industrial processes. In order to sustain either process, plants should accumulate proteins in sufficient quantities. Protein targeting to improve expression levels is a topic of major interest and additional studies on inducible expression are being pursued to enhance utility.

Currently, there is a major push to find and utilize renewable energy more efficiently. Plant starch and sugars are major sources of bioethanol, but their production from otherwise discarded lignocellulosic material hold out great promise for fuel production, enabling the real possibility of national fuel self-sufficiency. Plants are a renewable resource, and lower the LCA of processes since energy input into their production is substantially lower than other production systems. In addition, most products of plants can find use elsewhere for feed, silage, or biomass for renewable fuel production. The benefits of large-scale increases with little effort, lower costs, and potential to offset costs with downstream use of waste solely accrue to plants, making this a technology worth investing in.

## Bibliography

1. Brown T, Jones M, Powell W, Allaby R (2009) The complex origins of domesticated crops in the fertile crescent. *Trends Ecol Evol* 24:103–109
2. Hallauer A (2007) History, contribution, and future of quantitative genetics in plant breeding: lessons from maize. *Crop Sci* 47:54–519
3. Raskin I, Ribnicky D, Komarnytsky S, Ilic N, Poulev A, Borisjuk N, Brinker A, Moreno D, Ripoll C, Yakoby N (2002) Plants and human health in the twenty-first century. *Trends Biotechnol* 20:522–531
4. Giddings G (2001) Transgenic plants as protein factories. *Curr Opin Biotechnol* 12:450–454

5. Giddings G, Allison G, Brooks D, Carter A (2000) Transgenic plants as factories for biopharmaceuticals. *Nat Biotechnol* 18:1151–1155
6. Murphy D (2007) Improving containment strategies in biopharming. *Plant Biotechnol J* 5:555–569
7. Chapotin S, Wolt J (2007) Genetically modified crops for the bioeconomy: meeting public and regulatory expectations. *Transgenic Res* 16:675–688
8. Hansell A, Knorr-Held L, Best N, Schmid V, Aylin P (2003) Copd mortality trends 1950–1999 in England and Wales – did the 1956 clean air act make a detectable difference? *Epidemiology* 14:555
9. Hazilla M, Kopp R (1990) Social cost of environmental quality regulations: a general equilibrium analysis. *J Polit Econ* 98:853–873
10. Owens J (1997) Life-cycle assessment: Constraints on moving from inventory to impact assessment. *J Ind Ecol* 1:37–49
11. Harding K, Dennis J, Von Blottnitz H, Harrison S (2008) A life-cycle comparison between inorganic and biological catalysis for the production of biodiesel. *J Cleaner Prod* 16:1368–1378
12. van Beilen JB, Li Z (2002) Enzyme technology: an overview. *Curr Opin Biotechnol* 13:338–344
13. Kirk O, Borchert T, Fuglsang C (2002) Industrial enzyme applications. *Curr Opin Biotechnol* 13:345–351
14. Vajo Z, Fawcett J, Duckworth W (2001) Recombinant DNA technology in the treatment of diabetes: insulin analogs. *Endocr Rev* 22:706–717
15. Swartz J (2001) Advances in *Escherichia coli* production of therapeutic proteins. *Curr Opin Biotechnol* 12:195–201
16. Haki G, Rakshit S (2003) Developments in industrially important thermostable enzymes: a review. *Bioresour Technol* 89:17–34
17. Hatti-Kaul R, Törnvall U, Gustafsson L, Börjesson P (2007) Industrial biotechnology for the production of bio-based chemicals – a cradle-to-grave perspective. *Trends Biotechnol* 25:119–124
18. Seelig B, Szostak J (2007) Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. *Nature* 448:828–831
19. Jaeger K, Eggert T (2004) Enantioselective biocatalysis optimized by directed evolution. *Curr Opin Biotechnol* 15:305–313
20. Bornscheuer U, Pohl M (2001) Improved biocatalysts by directed evolution and rational protein design. *Curr Opin Chem Biol* 5:137–143
21. Cherry J, Fidantsef A (2003) Directed evolution of industrial enzymes: an update. *Curr Opin Biotechnol* 14:438–443
22. Jaeger K, Eggert T, Eipper A, Reetz M (2001) Directed evolution and the creation of enantioselective biocatalysts. *Appl Microbiol Biotechnol* 55:519–530
23. Louwrier A (1998) Industrial products-the return to carbohydrate-based industries. *Biotechnol Appl Biochem* 27:1–8
24. Zhao J, Li X, Qu Y (2006) Application of enzymes in producing bleached pulp from wheat straw. *Bioresour Technol* 97:1470–1476
25. Dyer T, Ragauskas A (2002) Developments in bleaching technology focus on reducing capital, operating costs. In: IPST Technical Paper Series Number 929 (Paper, P. a., Ed.)
26. Headon D, Walsh G (1994) The industrial production of enzymes. *Biotechnol Adv* 12:635–646
27. Chakrabarty A (1981) Microorganisms having multiple compatible degradative energy-generating plasmids and preparation thereof. US Patent No 4,259,444, Google Patents (Office, U. P., Ed.), General Electric Company, United States
28. Chakrabarty A, Kellogg S (1985) Bacteria capable of dissimilation of environmentally persistent chemical compounds. US Patent No 4,535,061. University of Illinois Foundation, United States
29. Sahdev S, Khattar S, Saini K (2008) Production of active eukaryotic proteins through bacterial expression systems: a review of the existing biotechnology strategies. *Mol Cell Biochem* 307:249–264
30. Twyman R, Schillberg S, Fischer R (2005) Transgenic plants in the biopharmaceutical market. *Expert Opin Emerg Drugs* 10:185–218
31. Dominguez A, Ferminan E, Sanchez M, Gonzalez F, Perez-Campo F, Garcia S, Herrero A, San Vicente A, Cabello J, Prado M (1998) Non-conventional yeasts as hosts for heterologous protein production. *Int Microbiol* 1:131–142
32. McAloon A, Taylor F, Yee W, Ibsen K, Wooley R (2000) Determining the cost of producing ethanol from corn starch and lignocellulosic feedstocks. National Renewable Energy Laboratory Report, NREL, Fort Collins, CO
33. Houdebine L (2009) Production of pharmaceutical proteins by transgenic animals. *Comp Immunol Microbiol Infect Dis* 32:107–121
34. Kretzmer G (2002) Industrial processes with animal cells. *Appl Microbiol Biotechnol* 59:135–142
35. Prusiner S, Scott M, DeArmond S, Cohen F (1998) Prion protein biology. *Cell* 93:337–348
36. Ragauskas A, Williams C, Davison B, Britovsek G, Cairney J, Eckert C, Frederick W Jr, Hallett J, Leak D, Liotta C (2006) The path forward for biofuels and biomaterials. *Science* 311:484
37. Sainz M (2009) Commercial cellulosic ethanol: the role of plant-expressed enzymes. *In Vitro Cell Dev Biol Plant* 45:314–329
38. Twyman RM, Stoger E, Schillberg S, Christou P, Fischer R (2003) Molecular farming in plants: host systems and expression technology. *Trends Biotechnol* 21:570–578
39. Howard JA, Hood E (2005) Bioindustrial and biopharmaceutical products produced in plants. *Adv Agron* 85:91–124
40. Borisjuk N, Borisjuk L, Logendra S, Petersen F, Gleba Y, Raskin I (1999) Production of recombinant proteins in plant root exudates. *Nat Biotechnol* 17:466–469
41. Hood E, Love R, Lane J, Bray J, Clough R, Pappu K, Drees C, Hood K, Yoon S, Ahmad A (2007) Subcellular targeting is a key condition for high-level accumulation of cellulase protein in transgenic maize seed. *Plant Biotechnol J* 5:709–719
42. Komarnytsky S, Borisjuk N, Borisjuk L, Alam M, Raskin I (2000) Production of recombinant proteins in tobacco guttation fluid. *Plant Physiol* 124:927–934

43. Evangelista R, Kusnadi A, Howard J, Nikolov Z (1998) Process and Economic evaluation of the extraction and purification of recombinant  $\beta$ -glucuronidase from transgenic corn. *Biotechnol Prog* 14:607–614
44. Doran P (2000) Foreign protein production in plant tissue cultures. *Curr Opin Biotechnol* 11:199–204
45. Fischer R, Stoger E, Schillberg S, Christou P, Twyman R (2004) Plant-based production of biopharmaceuticals. *Curr Opin Plant Biol* 7:152–158
46. Daniell H (2002) Molecular strategies for gene containment in transgenic crops. *Nat Biotechnol* 20:581–586
47. Lee D, Natesan E (2006) Evaluating genetic containment strategies for transgenic plants. *Trends Biotechnol* 24:109–114
48. Lin C, Fang J, Xu X, Zhao T, Cheng J, Tu J, Ye G, Shen Z (2008) A built-in strategy for containment of transgenic plants: creation of selectively terminable transgenic rice. *PLoS ONE* 3:e1818
49. Lienard D, Sourrouille C, Gomord V, Faye L (2007) Pharming and transgenic plants. *Biotechnol Annu Rev* 13:115–147
50. Daniell H, Streatfield S, Wycoff K (2001) Medical molecular farming: production of antibodies, biopharmaceuticals and edible vaccines in plants. *Trends Plant Sci* 6:219–226
51. Stoger E, Sack M, Fischer R, Christou P (2002) Plantibodies: applications, advantages and bottlenecks. *Curr Opin Biotechnol* 13:161–166
52. Orzáez D, Granell A, Blázquez M (2009) Manufacturing antibodies in the plant cell. *Biotechnol J* 4:1712–1724
53. Streatfield S, Howard J (2003) Plant production systems for vaccines. *Expert Rev Vaccin* 2:763–775
54. Lamphear B, Jilka J, Kesl L, Welter M, Howard J, Streatfield S (2004) A corn-based delivery system for animal vaccines: an oral transmissible gastroenteritis virus vaccine boosts lactogenic immunity in swine. *Vaccine* 22:2420–2424
55. Borkhardt B, Harholt J, Ulvskov P, Ahring B, Jørgensen B, Brinch-Pedersen H (2010) Autohydrolysis of plant xylans by apoplastic expression of thermophilic bacterial endoxylanases. *Plant Biotechnol J* 8:363–374
56. Woodard S, Mayor J, Bailey M, Barker D, Love R, Lane J, Delaney D, McComas-Wagner J, Mallubhotla H, Hood E (2003) Maize (*Zea mays*)-derived bovine trypsin: characterization of the first large-scale, commercial protein product from transgenic plants. *Biotechnol Appl Biochem* 38:123–130
57. Pinkerton T, Howard J, Wild J (2008) Genetically engineered resistance to organophosphate herbicides provides a new scoreable and selectable marker system for transgenic plants. *Mol Breed* 21:27–36
58. Bailey M, Woodard S, Callaway E, Beifuss K, Magallanes-Lundback M, Lane J, Horn M, Mallubhotla H, Delaney D, Ward M (2004) Improved recovery of active recombinant laccase from maize seed. *Appl Microbiol Biotechnol* 63:390–397
59. Clough RC, Pappu K, Thompson K, Beifuss K, Lane J, Delaney DE, Harkey R, Drees C, Howard JA, Hood EE (2006) Manganese peroxidase from the white-rot fungus *Phanerochaete chrysosporium* is enzymatically active and accumulates to high levels in transgenic maize seed. *Plant Biotechnol J* 4:53–62
60. Pen J, Molendijk L, Quax W, Sijmons P, van Ooyen A, van den Elzen P, Rietveld K, Hoekema A (1992) Production of active *Bacillus licheniformis* alpha-amylase in tobacco and its application in starch liquefaction. *Nat Biotechnol* 10:292–296
61. Pen J, Hoekema A, Sijmons P, Van Ooyen A, Rietveld K, Verwoerd T, Quax W (1991) Production of enzymes in seeds and their use. Gist-Brocades N.V. (Wateringseweg 1 P.O. Box 1, MA Delft, NL-2600, NL) Mogen International N.V. (Einsteinweg 97, CB Leiden, NL-2333, NL) European Patent
62. Czihal A, Conrad B, Buchner P, Brevis R, Farouk A, Manteuffel R, Adler K, Wobus U, Hofemeister J, Bäumllein H (1999) Gene farming in plants: expression of a heatstable *Bacillus* amylase in transgenic legume seeds. *J Plant Physiol* 155:183–189
63. Kumagai M, Donson J, Della-Cioppa G, Grill L (2000) Rapid, high-level expression of glycosylated rice-amylase in transfected plants by an RNA viral vector. *Gene* 245:169–174
64. Austin S, Bingham E, Mathews D, Shahan M, Will J, Burgess R (1995) Production and field performance of transgenic alfalfa (*Medicago sativa* L.) expressing alpha-amylase and manganese-dependent lignin peroxidase. *Euphytica* 85:381–393
65. Chiang C, Yeh F, Huang L, Tseng T, Chung M, Wang C, Lur H, Shaw J, Yu S (2005) Expression of a bi-functional and thermostable amylopullulanase in transgenic rice seeds leads to autohydrolysis and altered composition of starch. *Mol Breed* 15:125–143
66. Zhong G, Peterson D, Delaney D, Bailey M, Witcher D, Register Iii J, Bond D, Li C, Marshall L, Kulisek E (1999) Commercial production of aprotinin in transgenic maize seeds. *Mol Breed* 5:345–356
67. Pogue GP, Vojdani F, Palmer KE, Hiatt E, Hume S, Phelps J, Long L, Bohorova N, Kim D, Pauly M, Velasco J, Whaley K, Zeitlin L, Garger SJ, White E, Bai Y, Haydon H, Bratcher B (2010) Production of pharmaceutical-grade recombinant aprotinin and a monoclonal antibody product using plant-based transient expression systems. *Plant Biotechnol J* 8:1–17
68. Capell T, Bassie L, Christou P (2004) Modulation of the polyamine biosynthetic pathway in transgenic rice confers tolerance to drought stress. *Proc Natl Acad Sci USA* 101:9909–9914
69. Hood E, Witcher D, Maddock S, Meyer T, Baszczynski C, Bailey M, Flynn P, Register J, Marshall L, Bond D (1997) Commercial production of avidin from transgenic maize: characterization of transformant, production, processing, extraction and purification. *Mol Breed* 3:291–306
70. Samac D, Tesfaye M, Dornbusch M, Saruul P, Temple S (2004) A comparison of constitutive promoters for expression of transgenes in alfalfa (*Medicago sativa*). *Transgenic Res* 13:349–361
71. Mei C, Park S, Sabzikar R, Ransom C, Qi C, Sticklen M (2009) Green tissue-specific production of a microbial endo-cellulase in maize *Zea mays* L. endoplasmic-reticulum and mitochondria converts cellulose into fermentable sugars. *J Chem Technol Biotechnol* 84:689–695

72. Sun Y, Cheng J, Himmel M, Skory C, Adney W, Thomas S, Tisserat B, Nishimura Y, Yamamoto Y (2007) Expression and characterization of *Acidothermus cellulolyticus* E1 endoglucanase in transgenic duckweed *Lemna minor* 8627. *Bioresour Technol* 98:2866–2872
73. Dai Z, Hooker B, Quesenberry R, Thomas S (2005) Optimization of *Acidothermus cellulolyticus* endoglucanase (E1) production in transgenic tobacco plants by transcriptional, post-transcription and post-translational modification. *Transgenic Res* 14:627–643
74. Teymouri F, Alizadeh H, Laureano-Pérez L, Dale B, Sticklen M (2004) Effects of ammonia fiber explosion treatment on activity of endoglucanase from *Acidothermus cellulolyticus* in transgenic plant. *Appl Biochem Biotechnol* 116:1183–1191
75. Dai Z, Hooker B, Anderson D, Thomas S (2000) Expression of *Acidothermus cellulolyticus* endoglucanase E1 in transgenic tobacco: biochemical characteristics and physiological effects. *Transgenic Res* 9:43–54
76. Dai Z, Hooker B, Quesenberry R, Gao J (1999) Expression of *Trichoderma reesei* exo-cellobiohydrolase I in transgenic tobacco leaves and calli. *Appl Biochem Biotechnol* 79:689–699
77. Jensen L, Olsen O, Kops O, Wolf N, Thomsen K, Von Wettstein D (1996) Transgenic barley expressing a protein-engineered, thermostable (1,1,4)-beta-glucanase during germination. *Proc Natl Acad Sci USA* 93:3487–3491
78. Horvath H, Jensen L, Wong O, Kohl E, Ullrich S, Cochran J, Kannangara C, Von Wettstein D (2001) Stability of transgene expression, field performance and recombination breeding of transformed barley lines. *Theor Appl Genet* 102:1–11
79. Nuutila A, Ritola A, Skadsen R, Mannonen L, Kauppinen V (1999) Expression of fungal thermotolerant endo-1,4-beta-glucanase in transgenic barley seeds during germination. *Plant Mol Biol* 41:777–783
80. Ziegelhoffer T, Raasch J, Austin-Phillips S (2001) Dramatic effects of truncation and sub-cellular targeting on the accumulation of recombinant microbial cellulase in tobacco. *Mol Breed* 8:147–158
81. Dai Z, Hooker B, Anderson D, Thomas S (2000) Improved plant-based production of E1 endoglucanase using potato: expression optimization and tissue targeting. *Mol Breed* 6:277–285
82. Jin R, Richter S, Zhong R, Lamppa G (2003) Expression and import of an active cellulase from a thermophilic bacterium into the chloroplast both in vitro and in vivo. *Plant Mol Biol* 51:493–507
83. Kawazu T, Ohta T, Ito K, Shibata M, Kimura T, Sakka K, Ohmiya K (1996) Expression of a *Ruminococcus albus* cellulase gene in tobacco suspension cells. *J Ferment Bioeng* 82:205–209
84. Xue G, Patel M, Johnson J, Smyth D, Vickers C (2003) Selectable marker-free transgenic barley producing a high level of cellulase (1, 4-β-glucanase) in developing grains. *Plant Cell Rep* 21:1088–1094
85. Biswas G, Ransom C, Sticklen M (2006) Expression of biologically active *Acidothermus cellulolyticus* endoglucanase in transgenic maize plants. *Plant Sci* 171:617–623
86. Ransom C, Balan V, Biswas G, Dale B, Crockett E, Sticklen M (2007) Heterologous *Acidothermus cellulolyticus* 1, 4-beta-endoglucanase E1 produced within the corn biomass converts corn stover into glucose. *Appl Biochem Biotechnol* 137:207–219
87. Oraby H, Venkatesh B, Dale B, Ahmad R, Ransom C, Oehmke J, Sticklen M (2007) Enhanced conversion of plant biomass into glucose using transgenic rice-produced endoglucanase for cellulosic ethanol. *Transgenic Res* 16:739–749
88. Kawazu T, Sun J, Shibata M, Kimura T, Sakka K, Ohmiya K (1999) Expression of a bacterial endoglucanase gene in tobacco increases digestibility of its cell wall fibers. *J Biosci Bioeng* 88:421–425
89. Ziegler M, Thomas S, Danna K (2000) Accumulation of a thermostable endo-1, 4-β-D-glucanase in the apoplast of *Arabidopsis thaliana* leaves. *Mol Breed* 6:37–46
90. Ziegelhoffer T, Will J, Austin-Phillips S (1999) Expression of bacterial cellulase genes in transgenic alfalfa (*Medicago sativa* L.), potato (*Solanum tuberosum* L.) and tobacco (*Nicotiana tabacum* L.). *Mol Breed* 5:309–318
91. Yu L, Gray B, Rutzke C, Walker L, Wilson D, Hanson M (2007) Expression of thermostable microbial cellulases in the chloroplasts of nicotine-free tobacco. *J Biotechnol* 131:362–369
92. Kim S, Lee D, Choi I, Ahn S, Kim Y, Bae H (2010) *Arabidopsis thaliana* Rubisco small subunit transit peptide increases the accumulation of *Thermotoga maritima* endoglucanase Cel5A in chloroplasts of transgenic tobacco plants. *Transgenic Res* 19(3):489–497
93. Rooijen G, Glenn K, Shen, Y, Boothe J (2008) Commercial production of chymosin in plants, United States Patent Application 20080184394
94. Ralph J, Akiyama T, Kim H, Lu F, Schatz P, Marita J, Ralph S, Reddy M, Chen F, Dixon R (2006) Effects of coumarate 3-hydroxylase down-regulation on lignin structure. *J Biol Chem* 281:8843–8853
95. Montalvo-Rodriguez R, Haseltine C, Huess-LaRossa K, Clemente T, Soto J, Staswick P, Blum P (2000) Autohydrolysis of plant polysaccharides using transgenic hyperthermophilic enzymes. *Biotechnol Bioeng* 70:151–159
96. Meyer F, Smidansky E, Beecher B, Greene T, Giroux M (2004) The maize Sh2r6hs ADP-glucose pyrophosphorylase (AGP) large subunit confers enhanced AGP properties in transgenic wheat (*Triticum aestivum*). *Plant Sci* 167:899–911
97. Smidansky E, Clancy M, Meyer F, Lanning S, Blake N, Talbert L, Giroux M (2002) Enhanced ADP-glucose pyrophosphorylase activity in wheat endosperm increases seed yield. *Proc Natl Acad Sci U S A* 99:1724–1729
98. Reggi S, Marchetti S, Patti T, Amicis F, Cariati R, Bembi B, Fogher C (2005) Recombinant human acid β-glucosidase stored in tobacco seed is stable, active and taken up by human fibroblasts. *Plant Mol Biol* 57:101–113
99. Witcher D, Hood E, Peterson D, Bailey M, Bond D, Kusnadi A, Evangelista R, Nikolov Z, Wooge C, Mehig R (1998) Commercial production of β-glucuronidase (GUS): a model system for the production of proteins in plants. *Mol Breed* 4:301–312

100. Chan M, Chao Y, Yu S (1994) Novel gene expression system for plant cells based on induction of alpha-amylase promoter by carbohydrate starvation. *J Biol Chem* 269: 17635–17641
101. de Wilde C, Uzan E, Zhou Z, Kruus K, Andberg M, Buchert J, Record E, Asther M, Lomascolo A (2008) Transgenic rice as a novel production system for *Melanocarpus* and *Pycnopus* laccases. *Transgenic Res* 17:515–527
102. Gruber V, Berna P, Arnaud T, Bournat P, Clément C, Mison D, Olagnier B, Philippe L, Theisen M, Baudino S (2001) Large-scale production of a therapeutic protein in transgenic tobacco plants: effect of subcellular targeting on quality of a recombinant dog gastric lipase. *Mol Breed* 7:329–340
103. Zhong Q, Gu Z, Glatz C (2006) Extraction of recombinant dog gastric lipase from transgenic corn seed. *J Agric Food Chem* 54:8086–8092
104. Mokrzycki-Issartel N, Bouchon B, Farrer S, Berland P, Laparra H, Madelmont J, Theisen M (2003) A transient tobacco expression system coupled to MALDI-TOF-MS allows validation of the impact of differential targeting on structure and activity of a recombinant therapeutic glycoprotein produced in plants. *FEBS Lett* 552:170–176
105. Lagrimini L, Bradford S, Rothstein S (1990) Peroxidase-induced wilting in transgenic tobacco plants. *Plant Cell* 2:7–18, Online
106. Chen R, Xue G, Chen P, Yao B, Yang W, Ma Q, Fan Y, Zhao Z, Tarczynski M, Shi J (2008) Transgenic maize plants expressing a fungal phytase gene. *Transgenic Res* 17:633–643
107. Hamada A, Yamaguchi K, Harada M, Horiguchi K, Takahashi T, Honda H (2006) Recombinant, rice-produced yeast phytase shows the ability to hydrolyze phytate derived from seed-based feed, and extreme stability during ensilage treatment. *Biosci Biotechnol Biochem* 70:1524–1527
108. Brinch-Pedersen H, Hatzack F, Stoger E, Arcalis E, Pontopidan K, Holm P (2006) Heat-stable phytases in transgenic wheat (*Triticum aestivum* L.): deposition pattern, thermostability, and phytate hydrolysis. *J Agric Food Chem* 54:4624–4632
109. Zimmermann P, Zardi G, Lehmann M, Zeder C, Amrhein N, Frossard E, Bucher M (2003) Engineering the root-soil interface via targeted expression of a synthetic phytase gene in trichoblasts. *Plant Biotechnol J* 1:353–360
110. George T, Simpson R, Hadobas P, Richardson A (2005) Expression of a fungal phytase gene in *Nicotiana tabacum* improves phosphorus nutrition of plants grown in amended soils. *Plant Biotechnol J* 3:129–140
111. Brinch-Pedersen H, Olesen A, Rasmussen S, Holm P (2000) Generation of transgenic wheat (*Triticum aestivum* L.) for constitutive accumulation of an *Aspergillus* phytase. *Mol Breed* 6:195–206
112. Lucca P, Hurrell R, Potrykus I (2001) Genetic engineering approaches to improve the bioavailability and the level of iron in rice grains. *Theor Appl Genet* 102:392–397
113. Xiao K, Harrison M, Wang Z (2005) Transgenic expression of a novel *M. truncatula* phytase gene results in improved acquisition of organic phosphorus by Arabidopsis. *Planta* 222:27–36
114. Pen J, Verwoerd T, van Paridon P, Beudeker R, van den Elzen P, Geerse K, van der Klis J, Versteegh H, van Ooyen A, Hoekema A (1993) Phytase-containing transgenic seeds as a novel feed additive for improved phosphorus utilization. *Nat Biotechnol* 11:811–814
115. Hegeman C, Grabau E (2001) A novel phytase with sequence similarity to purple acid phosphatases is expressed in cotyledons of germinating soybean seedlings. *Plant Physiol* 126:1598–1608
116. Verwoerd T, Van Paridon P, Van Ooyen A, Van Lent J, Hoekema A, Pen J (1995) Stable accumulation of *Aspergillus niger* phytase in transgenic tobacco leaves. *Plant Physiol* 109:1199–1205
117. Ponstein A, Bade J, Verwoerd T, Molendijk L, Storms J, Beudeker R, Pen J (2002) Stable expression of phytase (phyA) in canola (*Brassica napus*) seeds: towards a commercial product. *Mol Breed* 10:31–44
118. Li J, Hegeman C, Hanlon R, Lacy G, Denbow D, Grabau E (1997) Secretion of active recombinant phytase from soybean cell-suspension cultures. *Plant Physiol* 114:1103–1111
119. Hong C, Cheng K, Tseng T, Wang C, Liu L, Yu S (2004) Production of two highly active bacterial phytases with broad pH optima in germinated transgenic rice seeds. *Transgenic Res* 13:29–39
120. Austin-Phillips S, Koegel R, Straub R, Cook M (1999) Animal feed compositions containing phytase derived from transgenic alfalfa and methods of use thereof. US Patent No 5,900,525, Wisconsin Alumni Research Foundation, Madison, Wisconsin, USA
121. Claparols M, Bassie L, Miro B, Del Duca S, Rodriguez-Montesinos J, Christou P, Serafini-Fracassini D, Capell T (2004) Transgenic rice as a vehicle for the production of the industrial enzyme transglutaminase. *Transgenic Res* 13:195–199
122. Yang P, Wang Y, Bai Y, Meng K, Luo H, Yuan T, Fan Y, Yao B (2007) Expression of xylanase with high specific activity from *Streptomyces olivaceoviridis* A1 in transgenic potato plants (*Solanum tuberosum* L.). *Biotechnol Lett* 29:659–667
123. Kimura T, Mizutani T, Tanaka T, Koyama T, Sakka K, Ohmiya K (2003) Molecular breeding of transgenic rice expressing a xylanase domain of the xynA gene from *Clostridium thermocellum*. *Appl Microbiol Biotechnol* 62:374–379
124. Liu J, Selinger L, Cheng K, Beauchemin K, Moloney M (1997) Plant seed oil-bodies as an immobilization matrix for a recombinant xylanase from the rumen fungus *Neocallimastix patriciarum*. *Mol Breed* 3:463–470
125. Patel M, Johnson J, Brettell R, Jacobsen J, Xue G (2000) Transgenic barley expressing a fungal xylanase gene in the endosperm of the developing grains. *Mol Breed* 6:113–124
126. Hyunjong B, Lee D, Hwang I (2006) Dual targeting of xylanase to chloroplasts and peroxisomes as a means to increase protein accumulation in plant cells. *J Exp Bot* 57:161–169

127. Bae H, Kim H, Kim Y (2008) Production of a recombinant xylanase in plants and its potential for pulp biobleaching applications. *Bioresour Technol* 99:3513–3519
128. Herbers K, Wilke I, Sonnewald U (1995) A thermostable xylanase from *Clostridium thermocellum* expressed at high levels in the apoplast of transgenic tobacco has no detrimental effects and is easily purified. *Biotechnology* 13:63–66
129. Herbers K, Flint H, Sonnewald U (1996) Apoplastic expression of the xylanase and (1–3, 1–4) glucanase domains of the xyn D gene from *Ruminococcus flavefaciens* leads to functional polypeptides in transgenic tobacco plants. *Mol Breed* 2:81–87
130. Estevez J, Cantero A, Reindl A, Reichler S, Leon P (2001) 1-Deoxy-D-xylulose-5-phosphate synthase, a limiting enzyme for plastidic isoprenoid biosynthesis in plants. *J Biol Chem* 276:22901–22909
131. Vain P (2007) Thirty years of plant transformation technology development. *Plant Biotechnol J* 5:221–229
132. Tzfira T, Citovsky V (2006) *Agrobacterium*-mediated genetic transformation of plants: biology and biotechnology. *Curr Opin Biotechnol* 17:147–154
133. Daniell H, Cohill P, Kumar S, Dufourmantel N, Dubald M (2007) Chloroplast genetic engineering. In: Daniell H, Chase C (eds) *Molecular biology and biotechnology of plant organelles*. Springer, Dordrecht
134. Marillonnet S, Thoeringer C, Kandzia R, Klimyuk V, Gleba Y (2005) Systemic *Agrobacterium tumefaciens*-mediated transfection of viral replicons for efficient transient expression in plants. *Nat Biotechnol* 23:718–723
135. Sainsbury F, Lomonosoff G (2008) Extremely high-level and rapid transient protein production in plants without the use of viral replication. *Plant Physiol* 148:1212–1218
136. Ueki S, Lacroix B, Krichevsky A, Lazarowitz S, Citovsky V (2008) Functional transient genetic transformation of *Arabidopsis* leaves by biolistic bombardment. *Nat Protoc* 4:71–77
137. Watson S (1988) Corn marketing, processing, and utilization. US 9,011,308 (NAL/USDA, A. U., Ed.)
138. Hood E, Kusnadi A, Nikolov Z, Howard J (1999) Molecular farming of industrial proteins from transgenic maize. In: Sahidi F, Kolodziejczyk P, Whitaker J, Munguia A, Fuller G (eds) *Chemicals via higher plant bioengineering*. Kluwer Academic/Plenum Publishers, New York, pp 127–148
139. Ramessar K, Sabalza M, Capell T, Christou P (2008) Maize plants: an ideal production platform for effective and safe molecular pharming. *Plant Sci* 174:409–419
140. Howard J, Hood E (2007) Methods for growing nonfood products in transgenic plants. *Crop Sci* 47:1255
141. Gepts P (2002) A comparison between crop domestication, classical plant breeding, and genetic engineering. *Crop Sci* 42:1780–1790
142. Gur A, Zamir D (2004) Unused natural variation can lift yield barriers in plant breeding. *PLoS Biol* 2:1610–1615
143. Dale P (1992) Spread of engineered genes to wild relatives. *Plant Physiol* 100:13
144. Delaney B, Astwood J, Cunny H, Conn R, Herouet-Guicheny C, MacIntosh S, Meyer L, Privalle L, Gao Y, Mattsson J (2008) Evaluation of protein safety in the context of agricultural biotechnology. *Food Chem Toxicol* 46:S71–S97
145. Nikolov Z, Hammes D (2002) Production of recombinant proteins from transgenic crops. In: Hood EE, Howard JA (eds) *Plants as factories for protein production*. Kluwer Academic Publishers, Dordrecht, pp 159–174
146. Horn M, Woodard S, Howard J (2004) Plant molecular farming: systems and products. *Plant Cell Rep* 22:711–720
147. Hellwig S, Drossard J, Twyman R, Fischer R (2004) Plant cell cultures for the production of recombinant proteins. *Nat Biotechnol* 22:1415–1422
148. Lam E, Benfey P, Gilmartin P, Fang R, Chua N (1989) Site-specific mutations alter *in vitro* factor binding and change promoter expression pattern in transgenic plants. *Proc Natl Acad Sci U S A* 86:7890–7894
149. Cai M, Wei J, Li X, Xu C, Wang S (2007) A rice promoter containing both novel positive and negative cis-elements for regulation of green tissue-specific gene expression in transgenic plants. *Plant Biotechnol J* 5:664–674
150. Belanger F, Kriz A (1991) Molecular basis for allelic polymorphism of the maize Globulin-1 gene. *Genetics* 129:863–872
151. Jones H, Sparks C (2009) Promoter sequences for defining transgene expression. In: Jones H, Shewry P (eds) *Transgenic wheat, barley and oats: production and characterization protocols*. Humana, Totowa
152. Streatfield S (2007) Approaches to achieve high-level heterologous protein production in plants. *Plant Biotechnol J* 5:2–15
153. Vance V, Vaucheret H (2001) RNA silencing in plants – defense and counterdefense. *Science* 292:2277–2280
154. Baulcombe D (2003) Overcoming and exploiting RNA silencing. In: Vasil IK (ed) 10th IAPTC&B congress. Kluwer Academic Publishers, Orlando, pp 49–58
155. Sticklen MB (2008) Plant genetic engineering for biofuel production: towards affordable cellulosic ethanol. *Nat Rev Genet* 9:433–443
156. Medrano G, Reidy M, Liu J, Ayala J, Dolan M, Cramer C (2009) Rapid system for evaluating bioproduction capacity of complex pharmaceutical proteins in plants. *Meth Mol Biol* 483:51–67
157. Van Rooijen G, Moloney M (1995) Structural requirements of oleosin domains for subcellular targeting to the oil body. *Plant Physiol* 109:1353–1361
158. Rigano M, Alvarez M, Pinkhasov J, Jin Y, Sala F, Arntzen C, Walmsley A (2004) Production of a fusion protein consisting of the enterotoxigenic *Escherichia coli* heat-labile toxin B subunit and a tuberculosis antigen in *Arabidopsis thaliana*. *Plant Cell Rep* 22:502–508
159. Garbarino J, Oosumi T, Belknap W (1995) Isolation of a polyubiquitin promoter and its expression in transgenic potato plants. *Plant Physiol* 109:1371–1378
160. Doran P (2006) Foreign protein degradation and instability in plants and plant tissue cultures. *Trends Biotechnol* 24:426–432

161. Kim K, Kwon S, Lee H, Hur Y, Bang J, Kwak S (2003) A novel oxidative stress-inducible peroxidase promoter from sweetpotato: molecular cloning and characterization in transgenic tobacco plants and cultured cells. *Plant Mol Biol* 51:831–838
162. Cramer C, Boothe J, Oishi K (1999) Transgenic plants for therapeutic proteins: linking upstream and downstream strategies. *Curr Top Microbiol Immunol* 240:95–118
163. Moloney M (2003) Oleosin partitioning technology for production of recombinant proteins in oil seeds. In: Vinci V, Parekh S (eds) *Handbook of industrial cell culture: mammalian, microbial, and plant cells*. Humana Press, Totowa
164. Merino ST, Cherry J (2007) Progress and challenges in enzyme development for biomass utilization. *Adv Biochem Eng Biotechnol* 108:95–120
165. Jorgensen H, Kristensen JB, Felby C (2007) Enzymatic conversion of lignocellulose into fermentable sugars: challenges and opportunities. *Biofuels, Bioprod Biorefin* 1:119–134, be
166. Howard JA, Nikolov Z, Hood EE (2011) Enzyme production systems for biomass conversion. In: Hood EE, Nelson P, Powell R (eds) *The Plant Biomass Conversion*. John Wiley & Sons Inc.
167. Hamelinck C, Hooijdonk G, Faaij A (2005) Ethanol from lignocellulosic biomass: techno-economic performance in short-, middle-and long-term. *Biomass Bioenergy* 28:384–410
168. Solomon B, Barnes J, Halvorsen K (2007) Grain and cellulosic ethanol: history, economics, and energy policy. *Biomass Bioenergy* 31:416–425
169. Howard JA, Hood EE (unpublished)
170. Sheehan J, Aden A, Paustian K, Killian K, Brenner J, Walsh M, Nelson R (2003) Energy and environmental aspects of using corn stover for fuel ethanol. *J Ind Ecol* 7:117–146
171. Groover AT, Fontana JR, Arroyo JM, Yordan C, McCombie WR, Martienssen RA (2003) Secretion trap tagging of secreted and membrane-spanning proteins using Arabidopsis gene traps. *Plant Physiol* 132:698–708
172. Cohen O, Kronman C, Velan B, Shafferman A (2004) Amino acid domains control the circulatory residence time of primate acetylcholinesterases in rhesus macaques (*Macaca mulatta*). *Biochem J* 378:117–128
173. Huang J, Sutliff T, Wu L, Nandi S, Bengel K, Terashima M, Ralston A, Drohan W, Huang N, Rodriguez R (2008) Expression and purification of functional human  $\alpha$ 1-antitrypsin from cultured plant cells. *Biotechnol Prog* 17:126–133
174. Ma J, Hein M (1995) Plant antibodies for immunotherapy. *Plant Physiol* 109:341–346
175. Garcia-Casado G, Sanchez-Monge R, Chrispeels M, Armentia A, Salcedo G, Gomez L (1996) Role of complex asparagine-linked glycans in the allergenicity of plant glycoproteins. *Glycobiology* 6:471–477
176. Cockburn A (2002) Assuring the safety of genetically modified (GM) foods: the importance of an holistic, integrative approach. *J Biotechnol* 98:79–106
177. Peterson R, Arntzen C (2004) On risk and plant-based biopharmaceuticals. *Trends Biotechnol* 22:64–66

## Plant Molecular Pharming, Pharmaceuticals for Human Health

ANDREAS SCHIERMEYER, STEFAN SCHILLBERG

Department of Plant Biotechnology, Fraunhofer IME, Aachen, Germany

### Article Outline

Glossary  
 Definition of the Subject and Its Importance  
 Introduction  
 Plant Transformation  
 Posttranslational Modification  
 Downstream Processing  
 Plant-Derived Vaccines  
 Plant-Made Pharmaceuticals in Advanced Development  
 Future Directions  
 Bibliography

### Glossary

***Agrobacterium tumefaciens*** Gram negative phytopathogenic soil bacterium belonging to the family Rhizobiaceae. *A. tumefaciens* naturally infects a variety of dicotyledonous plant species and induces the formation of tumors (galls) by transferring genes located within the T-DNA.

**Biopharmaceuticals** Drugs, produced using modern biotechnological methods such as recombinant DNA technology, comprising proteins and/or nucleic acids for therapeutic or in vivo diagnostic purposes.

**Epigenetic effects** Changes caused in gene expression patterns that are not due to changes in the nucleotide sequence of the DNA but due to nucleotide modifications by methylation or RNA-directed mechanisms.

**Glycosylation** The co- or posttranslational addition of carbohydrate moieties to polypeptides. The carbohydrates may be either N-linked (at asparagine or arginine side chains) or O-linked (at serine, threonine, tyrosine, hydroxylysine, or hydroxyproline side chains).

**Immunoglobulin (Ig)** Major component of the adaptive immune system with specific antibody activity.



Immunoglobulins are produced by lymphocytes and consist of four polypeptide chains: two identical heavy and two identical light chains. Immunoglobulin G (IgG) is the principal immunoglobulin of the plasma with a molecular weight of 150 kDa. Immunoglobulin A (IgA) is a dimeric 400 kDa molecule secreted by mucosal surfaces. Beside the heavy and the light chains an IgA molecule contains a joining chain and the secretory component. Immunoglobulin M (IgM) is produced early during the immune response. Secreted IgM molecules have a star-shaped pentameric structure.

**Molecular farming (also pharming)** The production of pharmaceutical or technical proteins in plants or animals.

**Monoclonal antibody** An immunoglobulin secreted by a single clone of antibody producing cells.

**Plastid** Plant organelle bound by a double membrane containing its own circular genome. Several types of plastids are known that originate from a common precursor the proplastid. The most prominent form is the chloroplast found in the green tissues.

**Posttranslational modification** Any modification that occurs once a polypeptide has been synthesized, for example, proteolytic processing, glycosylation, methylation, phosphorylation, and prenylation.

**Suspension culture** Technique for the cultivation of plant cells or tissues in liquid culture medium under aseptic conditions using shake flasks or fermentation vessels.

**T-DNA** Transfer DNA. Natural T-DNA is located on large tumor inducing (Ti) or hairy root inducing (Ri) plasmids, although for gene transfer to plants it has been moved onto a smaller, more convenient vector. The T-DNA is transferred to the plant cell with the help of a range of virulence factors, and once in the nucleus it may be stably integrated into the plant genome.

**Transformation** Transfer of genetic information into a cell by biological (*A. tumefaciens*) or physical means (e.g., gene gun).

**Transgene** A segment of DNA that is introduced into the genome of a host cell by transformation, and which integrates into the genome so that it is inherited like any other gene.

**Transient expression** The temporary expression of a transgene within a host cell without stable integration into the host genome.

**Vaccine** Preparation of immunogenic material that stimulates active immunity in the recipient without causing disease. Vaccines can be based on killed or attenuated microorganisms or isolated components of the disease causing agent (subunit vaccines).

**Viral vector** Genetic elements derived from viral genomes for the transient expression of transgenes in a host cell. Viral vectors have the ability to replicate autonomously in the host cell and might be able to infect distant cells (depending on the degree of engineering).

**Zinc-finger nucleases** Chimeric proteins consisting of a zinc-finger domain conferring sequence specific DNA binding and a nuclease domain for the introduction of a double-strand break at the target site.

### Definition of the Subject and Its Importance

The demand for therapeutic proteins has increased in recent years and modern biotechnological methods have, until recently, ensured the production of safe and effective biopharmaceuticals to meet this demand. Various production platforms are currently used in the pharmaceutical industry, most based on the fermentation of engineered pro- and eukaryotic microorganisms, insect cells, or mammalian cells. However, the growth of the market for biopharmaceuticals is predicted to outpace production capacity using these platforms in the next decade, so alternatives are necessary. Intact plants and plant cell cultures are suitable production systems for a wide range of therapeutic proteins and could help to fulfill the need for increased production capacity. The production of pharmaceutical proteins in plants began with a monoclonal antibody expressed in transgenic tobacco plants more than 20 years ago. Since then many different plant species have been genetically engineered to produce valuable pharmaceutical proteins using a variety of transformation methods. Major progress has been achieved in transformation and expression technology, the downstream processing of transgenic plant material and the adaptation of regulatory procedures to encompass the new production platforms, allowing the first plant-made pharmaceuticals to begin clinical trials.

## Introduction

Plant cells synthesize a vast array of secondary metabolites, many of which are already used as pharmaceuticals. Recombinant DNA technology combined with techniques for plant transformation and the regeneration of transgenic plants have allowed the pharmaceutical exploitation of plants to be extended to include the production of biopharmaceuticals such as subunit vaccines, antibodies, growth factors, cytokines, enzymes, and blood factors. In many cases, these products need to be purified from plant material and formulated in the same way as conventional biopharmaceuticals. However, many plants are “generally regarded as safe” for both topical and oral administration, so they are particularly suitable for the production of vaccines that can be delivered via the oral route or antibodies applied as topical microbicides, especially where such products are required on a large scale. This reflects the fact that plants can be grown inexpensively on an agricultural scale and that plant-derived pharmaceuticals for topical/oral administration would require only minimal processing. This could potentially bring the costs of production and distribution down to levels suitable for deployment in developing countries with limited financial resources and a poor medical infrastructure. This contribution describes the technologies that facilitate biopharmaceutical production in plants and plant cell cultures either through transient expression or stable transformation. It also discusses issues relating to posttranslational modification, extraction and purification, and regulatory compliance, focusing on those plant-derived pharmaceutical products that have advanced the furthest in the clinic. A compilation of selected technical achievements in plant molecular farming is provided in [Table 1](#).

## Plant Transformation

Pharmaceutical proteins can be produced in plants or plant cells either through transient expression or stable transformation. In the first case, the DNA encoding the protein is delivered into plant cells by *Agrobacterium tumefaciens* or a viral vector (or a combination of the two) but there is no integration of this DNA into the plant genome and the protein is synthesized for a few hours or days. In the second case, DNA delivered either by *A. tumefaciens* or a physical process such as particle

bombardment integrates into the plant genome and becomes a permanent locus, allowing long-term production of the recombinant protein and the transmission of the trait to subsequent generations.

Each method has advantages and disadvantages that need to be evaluated on a case by case basis for each pharmaceutical protein, depending on its intended use and the production scale. The production of an immunoglobulin via stable integration into the nuclear genome was first reported in 1989 when Hiatt and colleagues produced a monoclonal IgG-class antibody in tobacco leaves [1]. They introduced the coding sequences for the gamma heavy chain and kappa light chain of the immunoglobulin into independent tobacco lines and then crossed plants from each line to stack the transgenes in a single plant, which was able to produce the full antibody. The same strategy was used to produce a chimeric secretory (sIgA/G) antibody, although in this case four transgenes were required (encoding the kappa light chain, the chimeric alpha/gamma heavy chain, the joining chain, and the secretory component) and four lines were bred over two generations to generate the final production crop [4]. Later on the assembly of a chimeric secretory sIgA/G antibody could be achieved by simultaneous transformation of all four components in rice plants [20]. These examples clearly show how plant cells can produce even the most complex proteins and modify and assemble them into functional oligomeric structures (two different cell types are required in mammals to produce secretory IgA antibodies). Stable transformation of the nuclear genome enables the combination of several independent expression cassettes into a single transgenic line and also allows the introduction of a transgene from a laboratory cultivar into other varieties of the same species, to yield a germplasm that is particularly suited for a certain purpose. The latter strategy has been used to breed dent and sweet corn varieties that produce the HIV-specific antibody 2G12. The expression cassettes were initially introduced into a laboratory maize cultivar with little agronomic value and low yield [21]. Using conventional breeding transgenes can be transferred to a germplasm that is either inaccessible for direct transformation or that is particularly suited for the cultivation under certain climate conditions.

A drawback of stable nuclear transformation is the time needed to identify and establish a germplasm with

**Plant Molecular Farming, Pharmaceuticals for Human Health. Table 1** Selected research achievements in the field of plant molecular farming between 1989 and 2008

Year	Achievement	Reference
1989	Full-size antibody (mouse IgG) expression in tobacco	[1]
1990	First human protein (HSA) produced in tobacco and potato	[2]
1992	First vaccine candidate (HBsAg) expressed in tobacco	[3]
1995	First secretory antibody (sIgA) produced in tobacco	[4]
1995	Plant seed oilbodies as vehicles for protein production and purification	[5]
1996	Expression of a protein-based polymer in tobacco	[6]
1998	First clinical trial with a vaccine candidate produced in transgenic potato	[7]
1999	Transient expression of an antibody by <i>Agrobacterium</i> vacuum infiltration	[8]
1999	N-glycan analysis of a plant-produced antibody	[9]
2000	Human growth hormone produced in tobacco chloroplasts	[10]
2001	N-glycan modification by expression of a human $\beta$ -1,4 galactosyltransferase	[11]
2004	Knockout mutants of moss lacking plant-specific glycosylation	[12]
2004	Generation of <i>Arabidopsis</i> plants lacking plant-specific glycosylation	[13]
2005	<i>Agrobacterium</i> -mediated delivery of viral replicons	[14]
2006	Approval of a plant-made vaccine for veterinary medicine	[15]
2007	Production of glucocerebrosidase with terminal mannose residues	[16]
2008	Clinical phase I trial with plant-produced anti-idiotypic vaccines	[17]
2008	Engineering of a CMP-sialic acid pathway in plants	[18]
2008	Phase III clinical trial with plant-made glucocerebrosidase	[19]

the desired properties. A large number of primary transformants often need to be screened to identify plants showing high-level transgene expression. These lines then need to be analyzed at the molecular level to determine the number and arrangement of the transgenes. For breeding purposes single-copy integration events or multicopy single locus events with a regular transgene arrangement are preferred [22]. In contrast, multiple transgene copies with a complex integration pattern are likely to suffer from both transcriptional gene silencing (TGS) and posttranscriptional gene silencing (PTGS) [23]. The transgenic plants must also be analyzed for unwanted pleiotropic effects that could be caused by the transgene itself or by the changes that are brought about by its integration, since transgene integration following both *Agrobacterium*-mediated transformation and particle bombardment is a random process. Precise transgene integration at

a predefined locus can be achieved by homologous recombination (gene targeting) but this has not been possible for most plant species in the past due to its very low efficiency. A notable exception is the moss *Physcomitrella patens*, where transformation by homologous recombination is a straightforward and robust process [24]. In higher plants, efficient homologous recombination has become possible only recently with the use of zinc-finger endonucleases. These are engineered endonucleases containing zinc-finger motifs that bind precise DNA sequences and introduce double-strand breaks at the target site. This in turn stimulates DNA repair in the host, thereby facilitating homologous recombination. This has enabled the precise engineering of transgenic plants, although there have been no reports thus far concerning applications in molecular farming [25–28].

Another disadvantage of nuclear transgenic plants is that the target protein often accumulates at low levels, making them commercially unfeasible. This has been addressed by the use of plastid transformation, where the transgene is integrated in the circular chloroplast or chromoplast genome, typically by particle bombardment. Unlike nuclear transformation, homologous recombination is an efficient method for transgene integration into the plastid genome, allowing precise gene targeting. Every plastid contains several copies of the genome, and each plant cell contains many plastids [29]; therefore, it is possible in principle to generate plant cells containing several thousand transgene copies (and these are not subject to silencing because the TGS and PTGS mechanisms are not present in the plastid). To ensure the transgene is present in every copy of the plastid genome (the homoplasmic state), the primary transformants must undergo multiple rounds of selection and regeneration. This is generally achieved using the marker gene aminoglycoside 3''-adenylyltransferase, which confers resistance to the antibiotic spectinomycin [30]. The high transgene copy number and absence of silencing allows the accumulation of some target proteins to levels exceeding 10% of the total soluble protein (TSP) in the cell. Furthermore, since plastids are evolutionarily derived from bacteria, it is possible to express multiple genes as operons, producing polycistronic mRNA [31].

Many biopharmaceuticals are complex molecules that require several posttranslational processing steps to achieve a functional state. Plastids are equipped to form disulfide bridges, as demonstrated for the production of the human growth hormone somatotropin [10], and they can also assemble oligomers as demonstrated for the production of cholera toxin B-subunit (CTB), which assembled into functional GM1 ganglioside-binding pentamers [32]. Human serum albumin, which requires posttranslational removal of the N-terminal methionine residue, has also been produced successfully in plastids [33]. The enzyme methionine aminopeptidase cleaves off the initiating N-formylmethionine in plastids depending on the subsequent amino acid sequence context, and this must be considered when dealing with proteins that need to have intact N-termini. Another elegant approach for the production of proteins with a defined N-terminus is the expression of the target protein as an N-terminal

ubiquitin fusion protein. Endogenous ubiquitin-specific proteases then remove the ubiquitin moiety precisely, a strategy that has enabled the production of native somatotropin that carries an N-terminal phenylalanine residue [10].

Plastid transformation technique was limited to tobacco for many years but has recently expanded to incorporate certain crop species such as lettuce and tomato [34, 35]. Plastid transformation in crop plants opens new possibilities in the area of oral vaccines, where antigens are produced in the edible parts of plants and delivered via the oral route. A further advantage of plastid transformation is the biosafety aspect, since chloroplasts are inherited maternally in most species and are therefore not present in pollen [36]. However, there are also some limitations. Many biopharmaceuticals need to undergo co- and post-translational glycosylation in order to fold properly or in order to remain functional and stable, but this process does not occur in plastids. Certain target proteins also appear to be unstable or toxic when expressed in plastids, for example, the rotavirus coat protein VP6 and HIV p24 antigen undergo rapid degradation in the chloroplasts of older tobacco leaves, with significant accumulation only possible in the youngest leaves. A codon-optimized HIV p24 construct allowed homogenous expression but all the leaves turned yellow, and rearrangements were observed within the plastid DNA [37].

Transient expression allows more rapid production than stable transformation (nuclear or plastid). DNA encoding the pharmaceutical proteins is either included within a T-DNA cassette carried by an *A. tumefaciens* strain delivered into leaf tissue by vacuum infiltration [38, 39], or inserted into a viral vector that is used to infect the plant [40, 41]. Transient expression can be used for the rapid testing of expression constructs for subsequent stable transformation procedures or can be scaled up for use as production system in its own right. Most of the viral vectors are based on RNA viruses such as *Tobacco mosaic virus* (TMV), *Potato virus X* (PVX), and *Cowpea mosaic virus* (CPMV). These vectors have been used both to produce intact proteins and to produce chimeric virus particles that display peptide antigens on their surface. In such peptide display vectors, the target peptide is fused to the coat protein, and because each

particle has many copies of the coat protein (and hence the antigen), the particles are strongly immunogenic and can be used without additional adjuvants to provoke an immune response. The versatility of this approach has been demonstrated with an experimental rabies vaccine that induced a protective immune response in mice. Furthermore human volunteers who ingested spinach leaves infected with the recombinant virus mounted a humoral immune response [42]. Conventional viral vectors have a limited capacity, and larger transgenes tend to be truncated or eliminated altogether as the virus spreads. This has been addressed by developing a series of deconstructed viral vectors in which the coat protein gene is deleted to provide space for the transgene. In order to deliver these vectors to a maximum number of plant cells the entire recombinant virus genome is incorporated as a DNA copy into a T-DNA cassette and delivered by *A. tumefaciens* via vacuum infiltration [43]. Two similar systems have been developed, one described as the launch vector system [44] and the other as the magnification system [14]. They both exploit the ability of *A. tumefaciens* to infect a large range of plants, thereby extending the host range of the natural virus and using the efficient viral replication system to amplify the coding sequence of the target protein. In a proof of concept experiment, the accumulation of green fluorescent protein (GFP) peaked at 4 g kg<sup>-1</sup> fresh weight in *Nicotiana benthamiana* plants transformed by magnification [14]. The platform has been refined for the production of oligomeric proteins such as antibodies. Full-size IgG1 immunoglobulins have been produced successfully at levels of up to 0.5 g kg<sup>-1</sup> fresh weight, by introducing the light and heavy chain coding sequences into two independent noninterfering vectors based on TMV and PVX, respectively [45].

### Posttranslational Modification

Approximately 30% of all approved biopharmaceuticals contain N-linked glycans, so N-glycosylation is the most important posttranslational modification that needs to be taken into account when manufacturing recombinant biopharmaceuticals in plants. The mechanism of N-glycosylation is conserved between plants and mammals, beginning in the endoplasmic reticulum (ER) with the transfer of an oligosaccharide precursor molecule

onto asparagine residues within the sequence motif N-X-S/T (where X is any amino acid except proline). The precursor molecule is subsequently trimmed to yield a structure known as a high-mannose type glycan. The protein then passes into the Golgi apparatus where additional glycan modifications take place.

The final complex type glycan structures differ between plants and mammals (Fig. 1), in that plant glycoproteins contain core  $\beta$ 1,2-xylose and  $\alpha$ 1,3-fucose residues whereas mammalian glycoproteins contain  $\beta$ 1,4-galactose and terminal N-acetyl-neuraminic acid (sialic acid) residues [46]. Plant-specific glycosylation patterns have been found to induce an immune response upon injection in some mammalian species [47–49]. To prevent these immune responses several strategies have been explored to avoid the addition of plant-specific sugar residues to the glycan structure. One approach is the attachment of a C-terminal H/KDEL amino acid sequence motif to retain the target protein within the ER, thereby preventing exposure to the Golgi apparatus and the attachment of  $\beta$ 1,2-xylose and  $\alpha$ 1,3-fucose residues [50]. This strategy has been applied successfully in the production of a chimeric mouse/human IgG1 antibody against human chorionic gonadotropin [51]. An alternative strategy is the knockout or knockdown of the endogenous plant  $\beta$ 1,2-xylosyltransferase and  $\alpha$ 1,3-fucosyltransferase genes, which has been achieved in the moss *P. patens* by homologous recombination [12]. The double knockout mutant was used to express human erythropoietin devoid of plant-specific glycan structures [52]. In the model plant *Arabidopsis thaliana*, the  $\beta$ 1,2-xylosyltransferase and  $\alpha$ 1,3-fucosyltransferase genes have been knocked out by T-DNA insertional mutagenesis [13]. This plant line has been used to produce the anti-HIV antibody 2G12 with a humanized glycan structure [53]. In the duckweed *Lemna minor*, the human anti-CD30 antibody MDX-060 was produced without plant glycans by co-introducing inverted repeat transgenes matching the sequences of the  $\beta$ 1,2-xylosyltransferase and  $\alpha$ 1,3-fucosyltransferase genes, so that they were silenced by RNA interference (RNAi) [54]. This antibody also demonstrated stronger antibody-dependent cell-mediated cytotoxicity (ADCC) compared to its counterpart produced in Chinese hamster ovary (CHO) cells, reflecting the tenfold higher affinity of



processing. Proteolytic degradation has been observed irrespective of the subcellular localization of the target protein but the extracellular compartments (apoplast and culture medium) appear to be particularly rich in proteolytic enzymes [65–68]. This has been addressed using a number of strategies, including the co-expression of protease inhibitors [69–71] and the co-secretion of unrelated proteins that might act as bait for the proteases [72]. Although the proteases responsible for recombinant protein degradation have yet to be identified, there are indications that certain classes of proteases are involved (e.g., aspartic proteases, metalloproteases, and serine proteases) [66, 68]. Once the proteases responsible for recombinant protein degradation are known, knockout and knockdown strategies can be employed to reduce their abundance. However, proteases play a significant role in many aspects of plant development, stress responses, and pathogen defense, so their elimination may only be suitable for cell and tissue cultures that are grown under sterile and controlled conditions in the absence of pathogens.

### Downstream Processing

Most biopharmaceuticals are formulated as a purified product so the majority of biopharmaceuticals produced in plants must be extracted from plant tissue and then purified and formulated in the same way as conventional biopharmaceutical products. Regardless of the upstream production platform, downstream processing can account for up to 80% of the total manufacturing costs for a biopharmaceutical protein [73], but the first downstream processing steps are largely determined by the specific production host [74]. If the target protein is produced in plant suspension cells and secreted into the culture medium, the purification process can begin directly after the cells have been removed by filtration or centrifugation. If the protein is produced in an intact plant and/or if it accumulates inside the plant cell, it must be released by mechanical disruption in the presence of an appropriate extraction buffer and the extract must be clarified by filtration and/or centrifugation to remove debris, fibers, and other particulates. Aqueous two-phase partition is a useful initial purification step to remove polyphenols and cell debris from crude plant extracts [75, 76]. The removal of polyphenols is critical to prevent fouling of the

chromatography media used in subsequent purification steps [74, 77]. After clarification, the product may be captured from the feed if a suitable affinity chromatography resin is available, allowing a high level of purification in a single step. A wide range of natural affinity ligands and an increasing number of synthetic ligands (e.g., mercaptoethylpyridine, MEP HyperCel™) are available, particularly for the capture of antibodies [78]. After capture, polishing is usually achieved by the application of two or more orthogonal separation methods to achieve maximal purity and contaminant removal, for example, ion exchange, hydrophobic interaction, hydroxyapatite, and size exclusion chromatography [79]. If a capture step is not possible, these chromatography methods may be used for intermediate purification prior to polishing. Purification can be facilitated by engineering the physicochemical properties of the target protein through genetic fusions, although the fusion tag must be removed after purification to yield the authentic protein as a final product (a protease cleavage site adjacent to the tag can be used to achieve separation). A fusion with the hydrophobic plant protein oleosin enables enrichment of the target protein by floating centrifugation [80]. Alternatively, fusing the target protein to elastin-like polypeptides (ELPs) allows the target protein to be isolated by thermal phase transition [81].

Potential contaminants include macromolecules such as host cell proteins and nucleic acids, as well as small molecules such as secondary metabolites (e.g., nicotine). The removal of contaminants derived from plant-associated microbes must also be demonstrated, especially endotoxins from *A. tumefaciens* that can cause inflammatory responses in humans. The successful removal of these substances has recently been demonstrated for a monoclonal antibody that has been produced in *N. benthamiana* by magnification [82].

Biopharmaceuticals produced for human clinical trials must achieve certain quality criteria that are defined by current good manufacturing practice (cGMP). The regulations for biopharmaceutical manufacture are defined by the Food and Drug Administration (FDA) in the USA and by the European Medicines Agency (EMA) in the European Union. The production of pharmaceuticals using plant suspension cells is very similar in concept to conventional systems based on mammalian cells, but intact plants

cultivated in the greenhouse or in the open field are very different in concept and in practice. The FDA and EMEA have published guidance documents covering the production of pharmaceuticals in plants, and these might be refined further in the future [83]. Recently the Fraunhofer Institute for Molecular Biology and Applied Ecology in Aachen, Germany, obtained the first GMP license for the production of a plant-made pharmaceutical protein for clinical phase I trials in Europe. Based on this process the anti-HIV antibody 2G12 was produced in transgenic tobacco plants in the greenhouse.

### Plant-Derived Vaccines

Plants have been proposed as an alternative production platform for subunit vaccines, with the added advantage that storage tissues such as cereal grains and potato tubers may be used to keep the vaccine stable without the need for a cold chain and could even be used to administer oral vaccines without processing, thus reducing costs.

Antigens embedded in the plant cell matrix are protected against the acidic conditions in the stomach and are released gradually, allowing the induction of a mucosal immune response. Many antigens that could be used as vaccines in humans or farm animals have been produced in plants including the hepatitis B virus surface antigen [3, 84, 85], the Norwalk virus capsid protein [86–88], the *Escherichia coli* heat labile toxin [89–91], and the rabies glycoprotein [42, 92].

Phase I clinical trials in humans have been conducted for some oral vaccines. The coding sequence for the B-chain of the heat labile toxin from enterotoxigenic *E. coli* (LT-B) has been expressed in transgenic potato and maize. Human volunteers who ingested three 50-g or 100-g doses of peeled raw potato slices containing 0.5–1 mg of LT-B developed anti LT serum IgG antibodies (91%) and half of the vaccinees also produced secretory IgA antibodies in their stools [7]. Similarly human volunteers who ingested 2 g of defatted corn germ meal containing 1 mg of LT-B developed anti LT serum IgG and IgA and sIgA in their stools [93].

Transgenic potato tubers producing the major capsid protein of the Norwalk virus (which causes gastroenteritis) were fed to human volunteers in two or three 150-g doses containing ~500 µg of the protein. Higher levels of IgA antibody-secreting cells were observed in

more than 90% of the vaccinees, 20% produced serum IgG or IgM responses, and 30% produced anti-NVCP antibodies in their stools [88].

Two phase I clinical trials with plant-derived hepatitis B surface antigen (HBsAg) have been reported. In the first trial, transgenic lettuce (*Lactuca sativa*) was orally administered to seven seronegative individuals in three 200-g doses containing 0.5–1 µg of HBsAg within 5 weeks. After the third dose, all subjects developed serum anti-HB antibodies of up to 6.3 mIU/ml serum [94]. In the second trial, transgenic potato tubers were fed to individuals who had previously been vaccinated against hepatitis B. The vaccinees received two or three 100-g doses of raw peeled potatoes each containing ~800 µg HBsAg. Higher serum anti-HB titers were observed in 60% of the vaccinee group whereas there was no increase in the control group [95].

Recently, H1N1 and H5N1 influenza virus hemagglutinin (HA) have been transiently expressed in *N. benthamiana* [96]. Both proteins assembled into virus-like particles (VLPs) that budded from the plant plasma membrane, a desirable outcome because VLPs are polyvalent and therefore much more immunogenic than soluble subunit vaccines. Mice parenterally immunized with low doses (0.5 µg) of the VLPs were protected against a lethal challenge with influenza virus [96]. A phase I dose escalation study to assess the safety of a plant-derived H5 VLP in healthy volunteers was initiated in 2010 ([www.clinicaltrials.gov](http://www.clinicaltrials.gov); NCT00984945).

Although most vaccines are intended to induce an immune response against pathogens, their use is not limited to the prevention of infectious diseases. More recently, vaccines have been developed for the prevention or the treatment of certain types of cancer. A plant-derived vaccine for the treatment of non-Hodgkin's lymphoma (NHL) based on idiotype antibodies has been evaluated in a phase I trial [97]. NHL is a clonal disease of the B-cell lineage and the malignant cells carry specific immunoglobulins (idiotypes) on their surface. These idiotypes can be used to trigger a specific immune response. Because the idiotypes are different in each patient the vaccine has to be manufactured individually for each treated person. Currently the patient's tumor cells are expanded from a biopsy as human/mouse heteromyelomas. The monoclonal idiotype antibody is purified and coupled to an immunogenic carrier protein like keyhole limpet hemocyanin (KLH)



and injected into the patient usually together with granulocyte-macrophage colony stimulating factor (GM-CSF) as an adjuvant [98]. To shorten the time needed to manufacture the patient-specific vaccine a plant-based production system has been developed in which the coding sequences for the variable domains of the idiotype antibody are cloned from the patient's biopsy and inserted into a viral vector for the expression of a single chain antibody (scFv). *N. benthamiana* plants have been infected with such viruses allowing the scFv to be purified from leaves [99, 100]. Sixteen NHL patients who had previously received chemotherapy were treated with two different doses of the tobacco-derived idiotype vaccine either with or without a GM-CSF adjuvant [17]. Most of the treated patients developed a cellular immune response although only three patients developed a humoral immune response. Recently, the Bayer Group announced another phase I clinical trial with idiotype vaccines for NHL using the transient magnification technology developed by their subsidiary Icon Genetics. The ongoing study will enroll 30 patients with progressive or relapsing NHL. The patients will receive 12 injections over 16 months, each consisting of 1.0 mg of the personalized vaccine. The study is designed as a safety study to evaluate potential toxicity associated with the therapy but will also analyze the relevant immunological parameters of the patients. The final results of the study are expected in 2012 ([www.clinicaltrials.gov](http://www.clinicaltrials.gov); NCT01022255).

### Plant-Made Pharmaceuticals in Advanced Development

The most advanced plant-derived pharmaceutical in terms of clinical development is glucocerebrosidase manufactured in transgenic carrot suspension cells (prGCD, taliglucerase alpha, Uplyso™). Patients suffering from Gaucher disease, an inherited lysosomal storage disorder, cannot produce active glucocerebrosidase and need enzyme replacement therapy with recombinant glucocerebrosidase, which is currently produced in CHO cells (imiglucerase, Cerezyme™) [101]. This is currently one of the most expensive biopharmaceuticals, with an annual treatment cost of USD 200,000 per patient [102]. The purified recombinant imiglucerase needs to be processed enzymatically to expose terminal mannose residues that are required

for the efficient uptake of the enzyme into macrophages. The plant-derived counterpart, taliglucerase alpha, does not require these additional processing steps because it is targeted to the cell vacuole where the complex type N-glycans are trimmed to the paucimannose form exposing terminal mannose residues [16]. A phase III clinical trial with taliglucerase alpha was completed successfully in 2009 ([www.clinicaltrials.gov](http://www.clinicaltrials.gov), NCT00376168) and the substance currently awaits market approval from the FDA. Meanwhile patients can get access to taliglucerase alpha under an expanded access protocol [101].

Another plant-derived protein currently in clinical development is alpha-interferon (IFN- $\alpha$ 2b) for the treatment of chronic hepatitis C infections. IFN- $\alpha$ 2b has a low molecular weight (19 kDa, no glycan chains) and is therefore eliminated rapidly by renal filtration. Special formulations are required to increase its serum half-life, and this is achieved in the current formulation produced in *E. coli* (peginterferone alpha-2b; PEGIntron™), by attachment to polyethylene glycol. The plant-derived protein (Locteron™) is produced in duckweed and formulated in poly(ether-ester) microspheres to achieve controlled release over a defined period [103]. The plant-derived version has been tested successfully in a clinical phase I/II study ([www.clinicaltrials.gov](http://www.clinicaltrials.gov), NCT00593151) to establish its safety, tolerability, and efficacy compared to PEGIntron™. Currently two phase IIb clinical trials are underway to determine the optimal dose for the treatment of hepatitis C patients and its efficacy in combination with the antiviral compound ribavirin ([www.clinicaltrials.gov](http://www.clinicaltrials.gov), NCT00863239, NCT00953589).

The first recombinant biopharmaceutical on the market was insulin, which received regulatory approval in 1982. A large number of diabetes patients depend on insulin therapy so current demand for the protein exceeds 8 metric tons per year. This demand is currently met by production in *E. coli* and *Saccharomyces cerevisiae* [104, 105]. The successful production of active recombinant human insulin has also been demonstrated in oilseeds, where oleosin fusion can be used to facilitate purification. As stated above, oleosin is a hydrophobic protein component of the seed oilbodies and fusion proteins become concentrated in the oilbodies allowing their purification by floating centrifugation, enzymatic cleavage, and then standard polishing chromatography

methods [106]. For large-scale insulin production, the Canadian company SemBioSys Genetics Inc. uses safflower (*Carthamus tinctorius*) plants [107]. The company recently announced the successful completion of a phase I/II clinical trial with healthy volunteers demonstrating that the safflower-derived insulin (SBS-1000) is equivalent to the recombinant insulin currently on the market [108].

### Future Directions

Many biopharmaceutical proteins have been produced successfully in plants and plant cell cultures, clearly demonstrating the utility of plant-based production platforms. The demand for biopharmaceuticals is predicted to rise in the future based on the large number of ongoing clinical trials that involve recombinant pharmaceutical proteins, but current fermenter-based production platforms are already struggling to meet the demand. The enormous flexibility offered by different plant production systems and their specific advantages in terms of cost, safety, and scalability, means that plants could provide an alternative source for recombinant biopharmaceuticals when the capacity of current platforms is exhausted.

To become more competitive with the currently established protein production platforms an increase in productivity for the plant cell factories is mandatory. Therefore future research will aim to boost the protein accumulation levels in plant cells. To achieve this goal different strategies are pursued including, among others, gene amplification, high throughput screening for elite events, targeting the protein of interest to suited storage organelles or even to create them artificially, and to shield the target protein against proteolytic degradation. Systems biology approaches will help to identify cellular targets that can be subsequently engineered to further improve the plant cell as a protein production host. The engineering process itself will be more precise in the future by employing the newly developed techniques to facilitate homologous recombination within the nuclear genome.

Beside the quantity also the quality of the final product will be a major focus of future research and development. Especially the engineering of the glycosylation pattern bears a great potential to optimize the stability and efficacy of the biopharmaceutical product. A critical point with respect to the glycosylation pattern will be the

detailed understanding of plant-produced glycoproteins with the mammalian immune system. This will be an important prerequisite for tailoring plant-produced subunit vaccines and to avoid unintended side effects. With respect to oral vaccines reliable formulation and administration protocols have to be defined to ensure the anticipated outcome is achieved.

Further optimization of transient plant expression systems will help to address future needs for the delivery of vaccine, personalized medicine, and biopharmaceuticals for the treatment of orphan diseases. The rapid production cycle will also enable a timely reaction to emerging diseases, pandemics, or biohazards. However, unlike stable transgenic plant production systems, there are currently no specific regulatory guidelines for transient technologies, which are becoming a perceived barrier to their widespread use and commercialization. Therefore, the establishment and harmonization of international regulations for transient expression systems are needed to enable the commercial application of this promising technology.

### Bibliography

#### Primary Literature

1. Hiatt A, Cafferkey R, Bowdish K (1989) Production of antibodies in transgenic plants. *Nature* 342:76–78
2. Sijmons PC, Dekker BM, Schrammeijer B, Verwoerd TC, van den Elzen PJ, Hoekema A (1990) Production of correctly processed human serum albumin in transgenic plants. *Biotechnology* 8:217–221
3. Mason HS, Lam DM, Arntzen CJ (1992) Expression of hepatitis B surface antigen in transgenic plants. *Proc Natl Acad Sci USA* 89:11745–11749
4. Ma JK, Hiatt A, Hein M, Vine ND, Wang F, Stabila P, van Dolleweerd C, Mostov K, Lehner T (1995) Generation and assembly of secretory antibodies in plants. *Science* 268:716–719
5. van Rooijen GJ, Moloney MM (1995) Plant seed oil-bodies as carriers for foreign proteins. *Biotechnology* 13:72–77
6. Zhang XR, Urry DW, Daniell H (1996) Expression of an environmentally friendly synthetic protein-based polymer gene in transgenic tobacco plants. *Plant Cell Rep* 16:174–179
7. Tacket CO, Mason HS, Lososky G, Clements JD, Levine MM, Arntzen CJ (1998) Immunogenicity in humans of a recombinant bacterial antigen delivered in a transgenic potato. *Nat Med* 4:607–609
8. Vaquero C, Sack M, Chandler J, Drossard J, Schuster F, Monecke M, Schillberg S, Fischer R (1999) Transient expression of a tumor-specific single-chain fragment and a chimeric antibody in tobacco leaves. *Proc Natl Acad Sci USA* 96:11128–11133

9. Cabanes-Macheteau M, Fitchette-Laine AC, Loutelier-Bourhis C, Lange C, Vine ND, Ma JK, Lerouge P, Faye L (1999) N-Glycosylation of a mouse IgG expressed in transgenic tobacco plants. *Glycobiology* 9:365–372
10. Staub JM, Garcia B, Graves J, Hajdukiewicz PT, Hunter P, Nehra N, Paradkar V, Schlittler M, Carroll JA, Spatola L, Ward D, Ye G, Russell DA (2000) High-yield production of a human therapeutic protein in tobacco chloroplasts. *Nat Biotechnol* 18:333–338
11. Bakker H, Bardor M, Molthoff JW, Gomord V, Elbers I, Stevens LH, Jordi W, Lommen A, Faye L, Lerouge P, Bosch D (2001) Galactose-extended glycans of antibodies produced by transgenic plants. *Proc Natl Acad Sci USA* 98:2899–2904
12. Koprivova A, Stemmer C, Altmann F, Hoffmann A, Kopriva S, Gorr G, Reski R, Decker EL (2004) Targeted knockouts of *Physcomitrella* lacking plant-specific immunogenic N-glycans. *Plant Biotechnol J* 2:517–523
13. Strasser R, Altmann F, Mach L, Glossl J, Steinkellner H (2004) Generation of *Arabidopsis thaliana* plants with complex N-glycans lacking  $\beta$ 1, 2-linked xylose and core  $\alpha$ 1, 3-linked fucose. *FEBS Lett* 561:132–136
14. Marillonnet S, Thoeringer C, Kandzia R, Klimyuk V, Gleba Y (2005) Systemic *Agrobacterium tumefaciens*-mediated transfection of viral replicons for efficient transient expression in plants. *Nat Biotechnol* 23:718–723
15. Walsh G (2006) Biopharmaceutical benchmarks 2006. *Nat Biotechnol* 24:769–776
16. Shaaltiel Y, Bartfeld D, Hashmueli S, Baum G, Brill-Almon E, Galili G, Dym O, Boldin-Adamsky SA, Silman I, Sussman JL, Futerman AH, Aviezer D (2007) Production of glucocerebrosidase with terminal mannose glycans for enzyme replacement therapy of Gaucher's disease using a plant cell system. *Plant Biotechnol J* 5:579–590
17. McCormick AA, Reddy S, Reinl SJ, Cameron TI, Czerwinski DK, Vojdani F, Hanley KM, Garger SJ, White EL, Novak J, Barrett J, Holtz RB, Tuse D, Levy R (2008) Plant-produced idiotype vaccines for the treatment of non-Hodgkin's lymphoma: safety and immunogenicity in a phase I clinical study. *Proc Natl Acad Sci USA* 105:10131–10136
18. Castilho A, Pabst M, Leonard R, Veit C, Altmann F, Mach L, Glossl J, Strasser R, Steinkellner H (2008) Construction of a functional CMP-sialic acid biosynthesis pathway in *Arabidopsis*. *Plant Physiol* 147:331–339
19. Aviezer D, Almon-Brill E, Shaaltiel Y, Galili G, Chertkoff R, Hashmueli S, Galun E, Zimran A (2008) Novel enzyme replacement therapy for Gaucher disease: on-going phase III clinical trial with recombinant human glucocerebrosidase expressed in plant cells. *Mol Genet Metab* 93:S15–S15
20. Nicholson L, Gonzalez-Melendi P, van Dolleweerd C, Tuck H, Perrin Y, Ma JKC, Fischer R, Christou P, Stoger E (2005) A recombinant multimeric immunoglobulin expressed in rice shows assembly-dependent subcellular localization in endosperm cells. *Plant Biotechnol J* 3:115–127
21. Rademacher T, Sack M, Arcalis E, Stadlmann J, Balzer S, Altmann F, Quendler H, Stiegler G, Kunert R, Fischer R, Stoger E (2008) Recombinant antibody 2G12 produced in maize endosperm efficiently neutralizes HIV-1 and contains predominantly single-GlcNAc N-glycans. *Plant Biotechnol J* 6:189–201
22. Kohli A, Melendi PG, Abranches R, Capell T, Stoger E, Christou P (2006) The quest to understand the basis and mechanisms that control expression of introduced transgenes in crop plants. *Plant Signal Behav* 1:185–195
23. De Wilde C, Van Houdt H, De Buck S, Angenon G, De Jaeger G, Depicker A (2000) Plants as bioreactors for protein production: avoiding the problem of transgene silencing. *Plant Mol Biol* 43:347–359
24. Decker EL, Reski R (2004) The moss bioreactor. *Curr Opin Plant Biol* 7:166–170
25. Cai CQ, Doyon Y, Ainley WM, Miller JC, Dekelver RC, Moehle EA, Rock JM, Lee YL, Garrison R, Schulenberg L, Blue R, Worden A, Baker L, Faraji F, Zhang L, Holmes MC, Rebar EJ, Collingwood TN, Rubin-Wilson B, Gregory PD, Urnov FD, Petolino JF (2009) Targeted transgene integration in plant cells using designed zinc finger nucleases. *Plant Mol Biol* 69:699–709
26. Durai S, Mani M, Kandavelou K, Wu J, Porteus MH, Chandrasegaran S (2005) Zinc finger nucleases: custom-designed molecular scissors for genome engineering of plant and mammalian cells. *Nucleic Acids Res* 33:5978–5990
27. Porteus MH (2009) Plant biotechnology: zinc fingers on target. *Nature* 459:337–338
28. Tovkach A, Zeevi V, Tzfira T (2009) A toolbox and procedural notes for characterizing novel zinc finger nucleases for genome editing in plant cells. *Plant J* 57:747–757
29. Bendich AJ (1987) Why do chloroplasts and mitochondria contain so many copies of their genome? *Bioessays* 6:279–282
30. Svab Z, Maliga P (1993) High-frequency plastid transformation in tobacco by selection for a chimeric aadA gene. *Proc Natl Acad Sci USA* 90:913–917
31. De Cosa B, Moar W, Lee SB, Miller M, Daniell H (2001) Overexpression of the Bt cry2Aa2 operon in chloroplasts leads to formation of insecticidal crystals. *Nat Biotechnol* 19:71–74
32. Daniell H, Lee SB, Panchal T, Wiebe PO (2001) Expression of the native cholera toxin B subunit gene and assembly as functional oligomers in transgenic tobacco chloroplasts. *J Mol Biol* 311:1001–1009
33. Fernandez-San Millan A, Farran I, Molina A, Mingo-Castel AM, Veramendi J (2007) Expression of recombinant proteins lacking methionine as N-terminal amino acid in plastids: human serum albumin as a case study. *J Biotechnol* 127:593–604
34. Lelivelt CL, McCabe MS, Newell CA, Desnoo CB, van Dun KM, Birch-Machin I, Gray JC, Mills KH, Nugent JM (2005) Stable plastid transformation in lettuce (*Lactuca sativa* L.). *Plant Mol Biol* 58:763–774
35. Ruf S, Hermann M, Berger IJ, Carrer H, Bock R (2001) Stable genetic transformation of tomato plastids and expression of a foreign protein in fruit. *Nat Biotechnol* 19:870–875
36. Hagemann R (2002) Milestones in plastid genetics of higher plants. In: Esser K, Lüttge U, Beyschlag W, Hellwig F (eds)

- Progress in botany: genetics, physiology, ecology, vol 63. Springer, New York, pp 5–51
37. McCabe MS, Klaas M, Gonzalez-Rabade N, Poage M, Badillo-Corona JA, Zhou F, Karcher D, Bock R, Gray JC, Dix PJ (2008) Plastid transformation of high-biomass tobacco variety Maryland Mammoth for production of human immunodeficiency virus type 1 (HIV-1) p24 antigen. *Plant Biotechnol J* 6:914–929
  38. Fischer R, Vaquero-Martin C, Sack M, Drossard J, Emans N, Commandeur U (1999) Towards molecular farming in the future: transient protein expression in plants. *Biotechnol Appl Biochem* 30:113–116
  39. Sheludko YV (2008) Agrobacterium-mediated transient expression as an approach to production of recombinant proteins in plants. *Recent Pat Biotechnol* 2:198–208
  40. Lico C, Chen Q, Santi L (2008) Viral vectors for production of recombinant proteins in plants. *J Cell Physiol* 216:366–377
  41. Pogue GP, Lindbo JA, Garger SJ, Fitzmaurice WP (2002) Making an ally from an enemy: plant virology and the new agriculture. *Annu Rev Phytopathol* 40:45–74
  42. Yusibov V, Hooper D, Spitsin S, Fleysh N, Kean R, Mikheeva T, Deka D, Karasev A, Cox S, Randall J, Koprowski H (2002) Expression in plants and immunogenicity of plant virus-based experimental rabies vaccine. *Vaccine* 20:3155–3164
  43. Marillonnet S, Giritich A, Gils M, Kandzia R, Klimyuk V, Gleba Y (2004) In planta engineering of viral RNA replicons: efficient assembly by recombination of DNA modules delivered by Agrobacterium. *Proc Natl Acad Sci USA* 101:6852–6857
  44. Musyichuk K, Stephenson N, Bi H, Farrance CE, Orozovic G, Brodelius M, Brodelius P, Horsey A, Ugulava N, Shamloul AM, Mett V, Rabindran S, Streatfield SJ, Yusibov V (2007) A launch vector for the production of vaccine antigens in plants. *Influenza Other Respi Viruses* 1:19–25
  45. Giritich A, Marillonnet S, Engler C, van Eldik G, Botterman J, Klimyuk V, Gleba Y (2006) Rapid high-yield expression of full-size IgG antibodies in plants coinfecting with noncompeting viral vectors. *Proc Natl Acad Sci USA* 103:14701–14706
  46. Saint-Jore-Dupas C, Faye L, Gomord V (2007) From planta to pharma with glycosylation in the toolbox. *Trends Biotechnol* 25:317–323
  47. Bardor M, Faveeuw C, Fitchette AC, Gilbert D, Galas L, Trottein F, Faye L, Lerouge P (2003) Immunoreactivity in mammals of two typical plant glyco-epitopes, core  $\alpha(1, 3)$ -fucose and core xylose. *Glycobiology* 13:427–434
  48. Jin C, Altmann F, Strasser R, Mach L, Schahs M, Kunert R, Rademacher T, Glos J, Steinkellner H (2008) A plant-derived human monoclonal antibody induces an anti-carbohydrate immune response in rabbits. *Glycobiology* 18:235–241
  49. Bosch D, Schots A (2010) Plant glycans: friend or foe in vaccine development? *Expert Rev Vaccines* 9:835–842
  50. Tekoah Y, Ko K, Koprowski H, Harvey DJ, Wormald MR, Dwek RA, Rudd PM (2004) Controlled glycosylation of therapeutic antibodies in plants. *Arch Biochem Biophys* 426:266–278
  51. Sriraman R, Bardor M, Sack M, Vaquero C, Faye L, Fischer R, Finnern R, Lerouge P (2004) Recombinant anti-hCG antibodies retained in the endoplasmic reticulum of transformed plants lack core-xylose and core- $\alpha(1, 3)$ -fucose residues. *Plant Biotechnol J* 2:279–287
  52. Weise A, Altmann F, Rodriguez-Franco M, Sjoberg ER, Baumer W, Launhardt H, Kietzmann M, Gorr G (2007) High-level expression of secreted complex glycosylated recombinant human erythropoietin in the *Physcomitrella*  $\Delta$ -*fuc-t*  $\Delta$ -*xyl-t* mutant. *Plant Biotechnol J* 5:389–401
  53. Schähls M, Strasser R, Stadlmann J, Kunert R, Rademacher T, Steinkellner H (2007) Production of a monoclonal antibody in plants with a humanized N-glycosylation pattern. *Plant Biotechnol J* 5:657–663
  54. Cox KM, Sterling JD, Regan JT, Gasdaska JR, Frantz KK, Peele CG, Black A, Passmore D, Moldovan-Loomis C, Srinivasan M, Cuison S, Cardarelli PM, Dickey LF (2006) Glycan optimization of a human monoclonal antibody in the aquatic plant *Lemna minor*. *Nat Biotechnol* 24:1591–1597
  55. Gasdaska JR, Sterling JD, Regan JT, Cox KM, Sherwood SW, Dickey LF (2007) Expression of a glyco-optimized anti-CD20 antibody in the aquatic plant *Lemna* with enhanced ADCC activity. *Blood* 110:697a–697a
  56. Schuster M, Jost W, Mudde GC, Wiederkum S, Schwager C, Janzek E, Altmann F, Stadlmann J, Stemmer C, Gorr G (2007) In vivo glyco-engineered antibody with improved lytic potential produced by an innovative non-mammalian expression system. *Biotechnol J* 2:700–708
  57. Bakker H, Rouwendal GJ, Karnoup AS, Florack DE, Stoopen GM, Helsper JP, van Ree R, van Die I, Bosch D (2006) An antibody produced in tobacco expressing a hybrid  $\beta$ -1, 4-galactosyltransferase is essentially devoid of plant carbohydrate epitopes. *Proc Natl Acad Sci USA* 103:7577–7582
  58. Fujiyama K, Furukawa A, Katsura A, Misaki R, Omasa T, Seki T (2007) Production of mouse monoclonal antibody with galactose-extended sugar chain by suspension cultured tobacco BY2 cells expressing human  $\beta(1, 4)$ -galactosyltransferase. *Biochem Biophys Res Commun* 358:85–91
  59. Sourrouille C, Marquet-Blouin E, D'Aouist MA, Kiefer-Meyer MC, Seveno M, Pagny-Salehabadi S, Bardor M, Durambur G, Lerouge P, Vezina L, Gomord V (2008) Down-regulated expression of plant-specific glycoepitopes in alfalfa. *Plant Biotechnol J* 6:702–721
  60. Erbayraktar S, Grasso G, Sfacteria A, Xie QW, Coleman T, Kreilgaard M, Torup L, Sager T, Erbayraktar Z, Gokmen N, Yilmaz O, Ghezzi P, Villa P, Fratelli M, Casagrande S, Leist M, Helboe L, Gerwein J, Christensen S, Geist MA, Pedersen LO, Cerami-Hand C, Wuert JP, Cerami A, Brines M (2003) Asialoerythropoietin is a nonerythropoietic cytokine with broad neuroprotective activity in vivo. *Proc Natl Acad Sci USA* 100:6741–6746
  61. Wee EG, Sherrier DJ, Prime TA, Dupree P (1998) Targeting of active sialyltransferase to the plant Golgi apparatus. *Plant Cell* 10:1759–1768
  62. Castilho A, Strasser R, Stadlmann J, Grass J, Jez J, Gattinger P, Kunert R, Quendler H, Pabst M, Leonard R, Altmann F, Steinkellner H (2010) In planta protein sialylation through

- overexpression of the respective mammalian pathway. *J Biol Chem* 285:15923–15930
63. Karnoup AS, Turkelson V, Anderson WH (2005) O-linked glycosylation in maize-expressed human IgA1. *Glycobiology* 15:965–981
  64. Xu J, Tan L, Goodrum KJ, Kieliszewski MJ (2007) High-yields and extended serum half-life of human interferon  $\alpha$ 2b expressed in tobacco cells as arabinogalactan-protein fusions. *Biotechnol Bioeng* 97:997–1008
  65. De Muynck B, Navarre C, Nizet Y, Stadlmann J, Boutry M (2009) Different subcellular localization and glycosylation for a functional antibody expressed in *Nicotiana tabacum* plants and suspension cells. *Transgenic Res* 18:467–482
  66. Delannoy M, Alves G, Vertommen D, Ma J, Boutry M, Navarre C (2008) Identification of peptidases in *Nicotiana tabacum* leaf intercellular fluid. *Proteomics* 8:2285–2298
  67. Doran PM (2006) Foreign protein degradation and instability in plants and plant tissue cultures. *Trends Biotechnol* 24:426–432
  68. Schiermeyer A, Schinkel H, Apel S, Fischer R, Schillberg S (2005) Production of *Desmodus rotundus* salivary plasminogen activator  $\alpha$ 1 (DSPA $\alpha$ 1) in tobacco is hampered by proteolysis. *Biotechnol Bioeng* 89:848–858
  69. Kim TG, Lee HJ, Jang YS, Shin YJ, Kwon TH, Yang MS (2008) Co-expression of proteinase inhibitor enhances recombinant human granulocyte-macrophage colony stimulating factor production in transgenic rice cell suspension culture. *Protein Expr Purif* 61:117–121
  70. Komarnytsky S, Borisjuk N, Yakoby N, Garvey A, Raskin I (2006) Cosecretion of protease inhibitor stabilizes antibodies produced by plant roots. *Plant Physiol* 141:1185–1193
  71. Rivard D, Anguenot R, Brunelle F, Le VQ, Vezina L-P, Trepanier S, Michaud D (2006) An in-built proteinase inhibitor system for the protection of recombinant proteins recovered from transgenic plants. *Plant Biotechnol J* 4:359–368
  72. Baur A, Reski R, Gorr G (2005) Enhanced recovery of a secreted recombinant human growth factor using stabilizing additives and by co-expression of human serum albumin in the moss *Physcomitrella patens*. *Plant Biotechnol J* 3:331–340
  73. Roque AC, Lowe CR, Taipa MA (2004) Antibodies and genetically engineered related molecules: production and purification. *Biotechnol Prog* 20:639–654
  74. Menkhaus TJ, Bai Y, Zhang C, Nikolov ZL, Glatz CE (2004) Considerations for the recovery of recombinant proteins from plants. *Biotechnol Prog* 20:1001–1014
  75. Platis D, Drossard J, Fischer R, Ma JK, Labrou NE (2008) New downstream processing strategy for the purification of monoclonal antibodies from transgenic tobacco plants. *J Chromatogr A* 1211:80–89
  76. Platis D, Labrou NE (2009) Application of a PEG/salt aqueous two-phase partition system for the recovery of monoclonal antibodies from unclarified transgenic tobacco extract. *Biotechnol J* 4:1320–1327
  77. Bai Y, Glatz CE (2003) Bioprocess considerations for expanded-bed chromatography of crude canola extract: sample preparation and adsorbent reuse. *Biotechnol Bioeng* 81:775–782
  78. Low D, O'Leary R, Pujar NS (2007) Future of antibody purification. *J Chromatogr B* 848:48–63
  79. Drossard J (2004) Downstream processing of plant-derived recombinant therapeutic proteins. In: Fischer R, Schillberg S (eds) *Molecular farming: plant-made pharmaceuticals and technical proteins*. Wiley-VCH, Weinheim, Germany, pp 217–231
  80. Parmenter DL, Boothe JG, van Rooijen GJH, Yeung EC, Moloney MM (1995) Production of biologically active hirudin in plant seeds using oleosin partitioning. *Plant Mol Biol* 29:1167–1180
  81. Floss DM, Schallau K, Rose-John S, Conrad U, Scheller J (2010) Elastin-like polypeptides revolutionize recombinant protein expression and their biomedical application. *Trends Biotechnol* 28:37–45
  82. Pogue GP, Vojdani F, Palmer KE, Hiatt E, Hume S, Phelps J, Long L, Bohorova N, Kim D, Pauly M, Velasco J, Whaley K, Zeitlin L, Garger SJ, White E, Bai Y, Haydon H, Bratcher B (2010) Production of pharmaceutical-grade recombinant aprotinin and a monoclonal antibody product using plant-based transient expression systems. *Plant Biotechnol J* 8:638–654
  83. Spok A, Twyman RM, Fischer R, Ma JK, Sparrow PA (2008) Evolution of a regulatory framework for pharmaceuticals derived from genetically modified plants. *Trends Biotechnol* 26:506–517
  84. Kapusta J, Modelska A, Figlerowicz M, Pniewski T, Letellier M, Lisowa O, Yusibov V, Koprowski H, Plucienniczak A, Legocki AB (1999) A plant-derived edible vaccine against hepatitis B virus. *FASEB J* 13:1796–1799
  85. Kong Q, Richter L, Yang YF, Arntzen CJ, Mason HS, Thanavala Y (2001) Oral immunization with hepatitis B surface antigen expressed in transgenic plants. *Proc Natl Acad Sci USA* 98:11539–11544
  86. Mason HS, Ball JM, Shi JJ, Jiang X, Estes MK, Arntzen CJ (1996) Expression of Norwalk virus capsid protein in transgenic tobacco and potato and its oral immunogenicity in mice. *Proc Natl Acad Sci USA* 93:5335–5340
  87. Santi L, Batchelor L, Huang Z, Hjelm B, Kilbourne J, Arntzen CJ, Chen Q, Mason HS (2008) An efficient plant viral expression system generating orally immunogenic Norwalk virus-like particles. *Vaccine* 26:1846–1854
  88. Tacket CO, Mason HS, Losonsky G, Estes MK, Levine MM, Arntzen CJ (2000) Human immune responses to a novel norwalk virus vaccine delivered in transgenic potatoes. *J Infect Dis* 182:302–305
  89. Chikwamba R, Cunnick J, Hathaway D, McMurray J, Mason H, Wang K (2002) A functional antigen in a practical crop: LT-B producing maize protects mice against *Escherichia coli* heat labile enterotoxin (LT) and cholera toxin (CT). *Transgenic Res* 11:479–493
  90. Haq TA, Mason HS, Clements JD, Arntzen CJ (1995) Oral immunization with a recombinant bacterial antigen produced in transgenic plants. *Science* 268:714–716
  91. Rosales-Mendoza S, Soria-Guerra RE, Lopez-Revilla R, Moreno-Fierros L, Alpuche-Solis AG (2008) Ingestion of transgenic carrots expressing the *Escherichia coli* heat-labile enterotoxin B subunit protects mice against cholera toxin challenge. *Plant Cell Rep* 27:79–84

92. Ashraf S, Singh PK, Yadav DK, Shah Nawaz M, Mishra S, Sawant SV, Tuli R (2005) High level expression of surface glycoprotein of rabies virus in tobacco leaves and its immunoprotective activity in mice. *J Biotechnol* 119:1–14
93. Tacket CO, Pasetti MF, Edelman R, Howard JA, Streatfield S (2004) Immunogenicity of recombinant LT-B delivered orally to humans in transgenic corn. *Vaccine* 22:4385–4389
94. Kapusta J, Modelska A, Pniewski T, Figlerowicz M, Jankowski K, Lisowa O, Plucienniczak A, Koprowski H, Legocki AB (2001) Oral immunization of human with transgenic lettuce expressing hepatitis B surface antigen. *Adv Exp Med Biol* 495:299–303
95. Thanavala Y, Mahoney M, Pal S, Scott A, Richter L, Natarajan N, Goodwin P, Arntzen CJ, Mason HS (2005) Immunogenicity in humans of an edible vaccine for hepatitis B. *Proc Natl Acad Sci USA* 102:3378–3382
96. D'Aoust MA, Lavoie PO, Couture MM, Trepanier S, Guay JM, Dargis M, Mongrand S, Landry N, Ward BJ, Vezina LP (2008) Influenza virus-like particles produced by transient expression in *Nicotiana benthamiana* induce a protective immune response against a lethal viral challenge in mice. *Plant Biotechnol J* 6:930–940
97. Arntzen CJ (2008) Plant science. Using tobacco to treat cancer. *Science* 321:1052–1053
98. Sinha R, Shenoy PJ, Flowers CR (2008) Idiotype vaccine strategies for improving outcomes in follicular lymphoma. *Expert Opin Biol Ther* 8:1213–1223
99. McCormick AA, Kumagai MH, Hanley K, Turpen TH, Hakim I, Grill LK, Tuse D, Levy S, Levy R (1999) Rapid production of specific vaccines for lymphoma by expression of the tumor-derived single-chain Fv epitopes in tobacco plants. *Proc Natl Acad Sci USA* 96:703–708
100. McCormick AA, Reinl SJ, Cameron TI, Vojdani F, Fronfield M, Levy R, Tuse D (2003) Individualized human scFv vaccines produced in plants: humoral anti-idiotypic responses in vaccinated mice confirm relevance to the tumor Ig. *J Immunol Meth* 278:95–104
101. Hollak CE, vom Dahl S, Aerts JM, Belmatoug N, Bembi B, Cohen Y, Collin-Histed T, Deegan P, van Dussen L, Giraldo P, Mengel E, Michelakakis H, Manuel J, Hrebicek M, Parini R, Reinke J, di Rocco M, Pocovi M, Sa Miranda MC, Tytki-Szymanska A, Zimran A, Cox TM (2010) Force Majeure: therapeutic measures in response to restricted supply of imiglucerase (Cerezyme) for patients with Gaucher disease. *Blood Cells Mol Dis* 44:41–47
102. Kaiser J (2008) Is the drought over for pharming? *Science* 320:473–475
103. De Leede LG, Humphries JE, Bechet AC, Van Hoogdalem EJ, Verrijck R, Spencer DG (2008) Novel controlled-release *Lemna*-derived IFN- $\alpha$ 2b (Locteron): pharmacokinetics, pharmacodynamics, and tolerability in a phase I clinical trial. *J Interferon Cytokine Res* 28:113–122
104. Chance RE, Frank BH (1993) Research, development, production, and safety of biosynthetic human insulin. *Diab Care* 16(Suppl 3):133–142
105. Kjeldsen T (2000) Yeast secretory expression of insulin precursors. *Appl Microbiol Biotechnol* 54:277–286
106. Nykiforuk CL, Boothe JG, Murray EW, Keon RG, Goren HJ, Markley NA, Moloney MM (2006) Transgenic expression and recovery of biologically active recombinant human insulin from *Arabidopsis thaliana* seeds. *Plant Biotechnol J* 4:77–85
107. Moloney MM, Boothe J, Nykiforuk C, Kuhlman P, Pollock B (2008) Plant-based production of authentic recombinant human insulin: chemical, biochemical and preclinical evaluation. *FEBS J* 275:62–62
108. Boothe J, Nykiforuk C, Shen Y, Zaplachinski S, Szarka S, Kuhlman P, Murray E, Morck D, Moloney MM (2010) Seed-based expression systems for plant molecular farming. *Plant Biotechnol J* 8:588–606

### Books and Reviews

- Chen Q (2008) Expression and purification of pharmaceutical proteins in plants. *Biol Eng* 1:291–321
- Daniell H, Singh ND, Mason H, Streatfield SJ (2009) Plant-made vaccine antigens and biopharmaceuticals. *Trends Plant Sci* 14:669–679
- De Muynck B, Navarre C, Boutry M (2010) Production of antibodies in plants: status after twenty years. *Plant Biotechnol J* 8:529–563
- Faye L, Gomord V (eds) (2009) Recombinant proteins from plants: methods and protocols. Humana, New York
- Fischer R, Schillberg S (eds) (2004) Molecular farming: plant-made pharmaceuticals and technical proteins. Wiley-VCH, Weinheim, Germany
- Fischer R, Schillberg S, Twyman RM (2009) Molecular farming of antibodies in plants. In: Kirakosyan A, Kaufman PB (eds) Recent advances in plant biotechnology. Springer, Dordrecht, The Netherlands, pp 35–63
- Gomez E, Zoth SC, Berinstein A (2009) Plant-based vaccines for potential human application: a review. *Hum Vaccin* 5:738–744
- Hefferon KL (2009) Biopharmaceuticals in plants: toward the next generation of medicine. CRC, Boca Raton
- Hellwig S, Drossard J, Twyman RM, Fischer R (2004) Plant cell cultures for the production of recombinant proteins. *Nat Biotechnol* 22:1415–1422
- Karasev AV (ed) (2009) Plant-produced microbial vaccines. Springer, Berlin
- Kempken F, Jung C (eds) (2010) Genetic modification of plants: agriculture, horticulture and forestry. Springer, Berlin
- Knäblein J (2004) Biopharmaceuticals expressed in plants. In: Kayser O, Müller RH (eds) Pharmaceutical biotechnology. Wiley-VCH, Weinheim, Germany, pp 35–56
- Levine MM (ed) (2009) New generation vaccines, 4th edn. Informa Healthcare, New York
- Rader R (2007) Biopharmaceutical products in the US and European markets, 6th edn. Bioplan Associates, Rockville
- Rybicki EP (2010) Plant-made vaccines for humans and animals. *Plant Biotechnol J* 8:620–637

- Sharma AK, Sharma MK (2009) Plants as bioreactors: recent developments and emerging opportunities. *Biotechnol Adv* 27:811–832
- Shih SM, Doran PM (2009) Foreign protein production using plant cell and organ cultures: advantages and limitations. *Biotechnol Adv* 27:1036–1042
- Tiwari S, Verma PC, Singh PK, Tuli R (2009) Plants as bioreactors for the production of vaccine antigens. *Biotechnol Adv* 27:449–467
- Twyman RM, Schillberg S, Fischer R (2005) Transgenic plants in the biopharmaceutical market. *Expert Opin Emerg Drugs* 10:185–218
- Walsh G (2009) Market development of biopharmaceuticals. In: Engelhard M, Hagen K, Boysen M (eds) *Genetic engineering in livestock: new applications and interdisciplinary perspectives*. Springer, Berlin, pp 69–89
- Walsh G (ed) (2009) *Post-translational modification of protein biopharmaceuticals*. Wiley-VCH, Weinheim, Germany

## Plant Molecular Pharming, Veterinary Applications

DOREEN M. FLOSS<sup>1</sup>, UDO CONRAD<sup>2</sup>

<sup>1</sup>Institute of Biochemistry and Molecular Biology II, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

<sup>2</sup>Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Phytoantibodies, Gatersleben, Germany

### Article Outline

Glossary

Definition of the Subject

Introduction

Plant-Based Expression Systems

Edible Vaccines and Purification

Future Directions

Acknowledgments

Bibliography

### Glossary

**Companion animal** Animal kept for companionship and enjoyment (household animal).

**Edible vaccines** Antigenic proteins, which are produced in organs of transgenic plants (e.g., fruits, tubers) and can be directly administered to humans or animals without any purification procedure.

**ELP** Elastin-like polypeptide containing the hydrophobic amino acids valine, proline, glycine, and guest residues, which shows temperature-dependent, reversible self-aggregation.

**ELPylation** Genetic C- or N-terminal target protein fusion to elastin-like polypeptides.

**Homoplasmy** Presence of the transgene in all copies of chloroplast DNA.

**Livestock animal** Domesticated animal raised in an agricultural setting to produce, e.g., food and fiber.

**Molecular pharming** The large-scale production and purification of pharmaceutical proteins in plants.

**Plant-based expression** Process by which information from a transgene is used in the synthesis of a functional protein *in planta*. Different plant-based expression systems are suitable (e.g., transgenic plants, transient expression, and plant cell cultures).

**Plantibodies** Antibody or antibody derivative produced in genetically engineered plants.

**Transgenic plants** Genetically engineered plants generated by the biolistic method (particle gun) or by *Agrobacterium tumefaciens* mediated transformation. The introduced transgene, which does not occur naturally in the plant, is transferred to the offspring.

**Transient expression** Expression of transgenes for a short period of time. In the context of plant-based expression infiltration of recombinant *Agrobacteria* (Agro-infiltration) or the use of plant viral vectors are the methods of choice to produce a desired protein *in planta*.

**Transplastomic plants** Introduced transgene is targeted to the chloroplast genome using particle bombardment or other physical DNA delivery techniques.

**Zoonotic diseases** Infectious disease that can be transmitted from wild and domestic animals to humans.

### Definition of the Subject

“Molecular Pharming” refers to the large-scale production and purification of pharmaceutical proteins in plants or plant-based expression systems. Since the successful expression of complete antibodies in transgenic plants in 1989 and the first report of plant-based

vaccine production in 1992, a large number of different vaccines, antibodies, as well as antibody fragments have been produced in plants for medical or veterinary purposes. However, only two plant-produced vaccine-related products have gone all the way through the production and regulatory hurdles, and only one, a plant-made single-chain variable fragment (scFv), is used in the production of a recombinant Hepatitis B Virus (HBV) vaccine in Cuba. Edible vaccines and novel methods of downstream processing such as “ELPylation” have been developed over the past years to facilitate the development of recombinant protein-based therapeutics.

## Introduction

Plant-based expression systems possess advantages over conventional eukaryotic expression systems (yeast, insect cells, and mammalian cells), e.g., the ability to obtain complex, correctly folded, and posttranslationally modified proteins [50]. They compete favorably with mammalian cells for the production of vaccines and antibodies because of distinct advantages over conventional systems including cost, safety, and scalability [57]. However, the cost of downstream processing (protein extraction, protein recovery, and protein purification) for recombinant expression systems in general are approximately the same and can represent over 80% of the overall processing costs [30] with the majority of such costs attributed to chromatography and associated materials, labor, and capital equipment [57]. Savings in the upstream components (no need for expensive fermenters, special culture media, and skilled workers) are some of the major benefits for the production of pharmaceutical proteins in plants. Costs of goods sold (COGS) from mammalian cell culture are estimated to be \$300 per gram therapeutic protein, whereas the raw material costs for 1 g recombinant protein from plants are in the order of \$0.10–\$1 (depending on the expression level; [33]). The main technical bottleneck limiting the commercial production of pharmaceuticals in plants is the high cost and inefficiency of downstream processing including purification [34].

One-third of the approved biopharmaceuticals are glycoproteins [56] and the activity of antibodies, blood factors, and interferons is dependent on their

glycosylation pattern. Accordingly, biopharmaceuticals are often produced in heterologous expression systems with glycosylation capabilities. Plant-specific glycosylation differs from mammalian glycosylation (for review see [26]) and this aspect explains the major limitation for the use of plant-made pharmaceuticals in therapy. Recently, progress toward the humanization of protein N-glycosylation in plant cells has been made, which focused on the targeted expression of therapeutic proteins, the knock-out of plant-specific N-glycan-processing genes, and/or the introduction of the enzymatic machinery catalyzing the synthesis, transport, and addition of mammalian sugar residues (for review see [27]).

With the development of “edible” vaccines, which can be orally administered in the form of a transgenic fruit or vegetable expressing the appropriate antigen without any prior processing, low-cost production systems and effective delivery systems are expected [40]. One of the easiest ways to get vaccinated against a disease might be eating a bite of banana, full of the virus proteins, as it was contemplated by researchers at the Boyce Thompson Institute for Plant Research at Cornell University in 1997 [25]. In reality, this anticipated development did not occur. Major problems of this technology are low yields, weak antigenicity of plant-produced vaccines and the lack of buy-in by governments and pharmaceutical companies [43].

In this chapter, the historical development of plant-produced vaccines and antibodies, so-called plantibodies, and the development of different stable plant production systems including downstream processing with a specific focus on the progress of animal therapeutics will be discussed.

## Plant-Based Expression Systems

Since the first report of plant-based antibody production [32], different formats have been generated ranging from single variable heavy-chain domain (VHH) antibodies [4] and single-chain molecules (scFvs; [3, 23]) to Fab fragments [10], and complete immunoglobulins [35]. Despite substantial progress in the production of antibodies in plants for human health (for review see [9]), their application to the veterinary field is rather limited (for review see [18]). The development of passive immunization commenced in 1890 with the identification of serum therapy by Emil Behring and



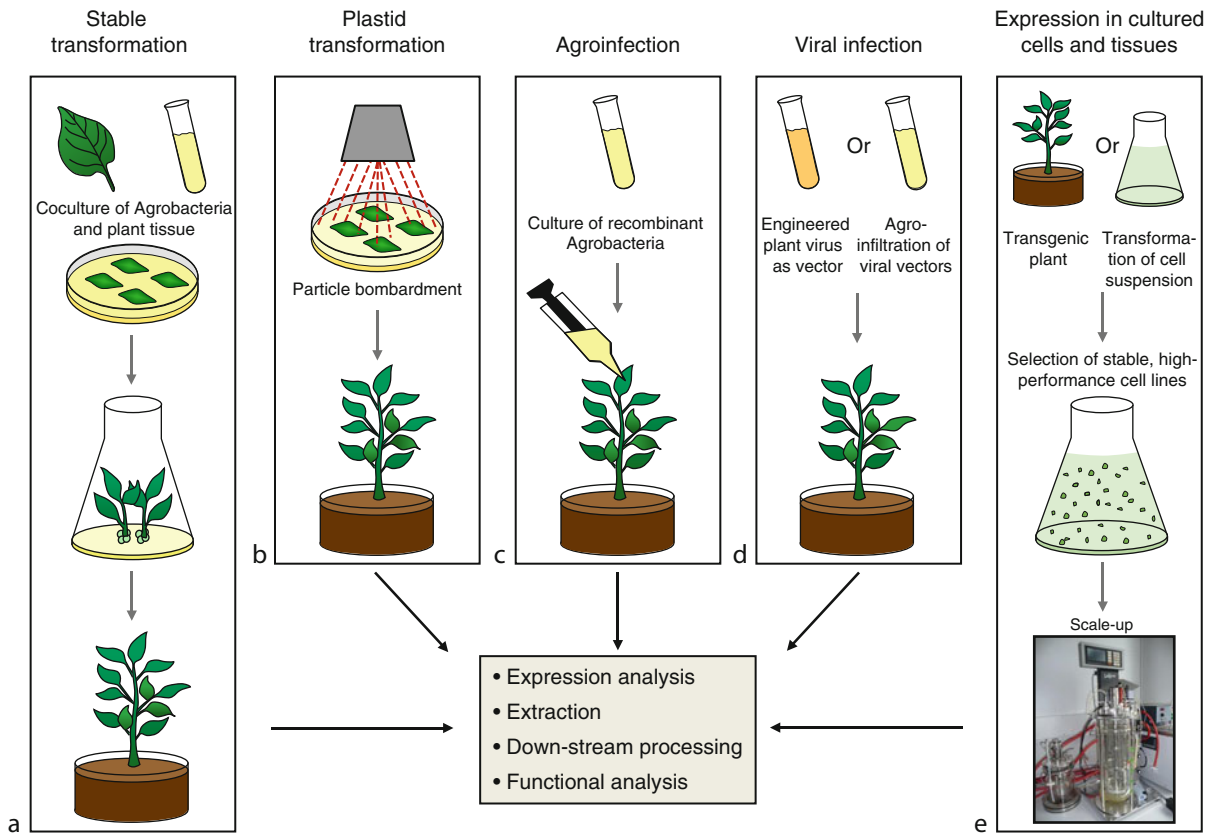
Shibasaburo Kitasato. They identified substances in the blood they called antibodies which were responsible for the immunity against diphtheria and tetanus toxins. Furthermore, the researchers were able to transfer immunity to immunologically naïve animals by injecting serum of animals treated with nonlethal doses of a crude toxin preparation [2]. At the same time, Paul Ehrlich discovered that antibodies can act as so-called magic bullets for the targeting of cancer cells [12]. A century later, the structures of antibodies and the sequences coding for the immunoglobulin chains were elucidated and mouse hybridomas provided highly specific monoclonal antibodies for therapeutic applications [36]. The development of innovative recombinant DNA technologies greatly enhanced the clinical efficiency and safety of mouse-derived monoclonal antibodies. The ability to generate large antibody libraries, and the simplified antibody backbone of a single-chain antibody made antibody phage display a powerful tool for the development of new therapeutic agents (for reviews see [3]).

Production shortfalls and high costs are major incentives for further development of alternative antibody production technologies with a focus on active immunization (vaccination). The defining event for the development of “vaccinology” (from the Latin “vacca,” meaning cow) dates back more than 200 years. At that time the smallpox vaccine was discovered by Edward Jenner. He inoculated humans with a less virulent, but antigenic related, Cowpox Virus to confer protection against the related human Smallpox Virus. Criteria for the development of veterinary vaccines are different depending on the particular target animals (for review see [38]). Health and welfare of the individual animal are the primary concerns for companion animal vaccines and thus are comparable to those for humans. In contrast, livestock animal vaccines should be inexpensive, prevent, and control infectious diseases of animals used as food to reduce or eliminate health risks to consumers. In some cases, these vaccines are further used to improve the productivity of livestock. To combat zoonotic diseases which are transmittable to humans, e.g., rabies, vaccination of wildlife animals is the method of choice. Furthermore, veterinary vaccines have a significant impact on public health due to the reduction in the administration of veterinary pharmaceuticals such as hormones.

The pioneering work for the expression of vaccines in plants was described in a patent by Curtiss and Cardineau in 1990. They reported the production of the *Streptococcus mutans* surface protein antigen A (SpaA) in transgenic tobacco plants [6]. Subsequently, Mason and co-workers succeeded in expressing the hepatitis B surface antigen in tobacco [37]. In 1993, Usha and co-workers expressed a peptide representing an epitope of the VP1 envelope protein of the Foot-and-Mouth-Disease Virus (FMDV) on the surface of a plant virus particle [55]. This study was the first report of a plant-derived veterinary vaccine. Following this pioneering work, various veterinary candidate vaccines have been produced in a variety of plant species using different expression systems, and they have proven to elicit humoral and mucosal immune responses against toxins, viruses, bacteria, and parasitic pathogens (for reviews see [18, 29, 44, 52, 57]). There are still no plant-derived veterinary vaccines on the market; however, one major step was made at the beginning of 2006 by Dow AgroSciences (DAS, Indianapolis, USA). Their plant cell-expressed vaccine against the Newcastle Disease Virus (NDV), produced in a suspension-cultured tobacco cell line, has gained regulatory approval by the US Department of Agriculture (USDA) Center for Veterinary Biologics – the final authority for veterinary vaccines in the USA [48]. Regrettably, this vaccine has not been introduced to the market. Dow AgroSciences apparently wished to demonstrate that their Concert™ Plant-Cell-Produced system was useful for the production of safe and effective vaccines, fulfilling the approval requirements of the regulatory system [43]. A year prior to the approval of the DAS vaccine, a plant-made scFv, used in the production of a recombinant Hepatitis B Virus (HBV) vaccine in Cuba [41], progressed through the regulatory system and was commercialized. These are the two plant-produced vaccine-related products which have gone through the production and regulatory hurdles, despite nearly 20 years of plant-derived vaccines [43].

Four plant-based expression systems have been developed thus far (Fig. 1):

- Expression in stably transformed transgenic plants including tissue-specific expression (e.g., in seeds or tubers)
- Expression in transplastomic plants



### Plant Molecular Pharming, Veterinary Applications. Figure 1

Plant-based expression systems for pharmaceutical proteins. **(a)** Transgenic plants derived by stable transformation, either using *Agrobacterium*-mediated gene transfer [54] or biolistic transformation [1], represent a stable and cheap source for the large-scale production of recombinant proteins. The transgene is genetically fixed and transferred into the next generation. However, the development as well as the selection of a stable transgenic line can take many months. Recombinant proteins may be expressed in the cytoplasm or be localized in other cellular compartments (nucleus, mitochondria, chloroplasts, vacuole, endoplasmic reticulum, or apoplast), or can be produced in different plant tissues (leaves, seeds). **(b)** Transplastomic plants obtained by using particle bombardment often have high yields of the recombinant proteins. However, the system is often not suitable for glycosylated or secreted proteins but this barrier may be overcome soon [27]. **(c)** *Agrobacterium tumefaciens*-mediated transient expression is the standard method for determining if a transgene is expressed *in planta*. Here, a suspension of bacteria is directly injected into the intercellular space of plant leaves either by using a syringe or vacuum. **(d)** Viral vectors can be used for the expression of foreign proteins or of chimeric coat proteins in plants. Two different methods can be used for the delivery of the viral genomes into the plant, either engineered plant viruses (e.g., Tobacco Mosaic Virus) or recombinant *Agrobacterium*. **(e)** The application of plant cell culture for the production of recombinant proteins is focused on a small number of cell lines, e.g., the tobacco line Bright Yellow-2 (BY-2). Furthermore, transgenic cell lines can be established either from a transgenic plant or by the transformation of cell suspensions either by *Agrobacterium* or particle bombardment. After selection of stable, high-performance cell lines, the recombinant proteins can be produced in bioreactors under “good manufacturing practice” (GMP)

- Transient expression in tobacco leaves (*Nicotiana tabacum*, *N. benthamiana*) using either plant viruses, *Agrobacterium tumefaciens*, or both to facilitate high accumulation of vaccines and/or antibodies
- Expression in cultured plant cells and tissues, and lower plants including duck weed and mosses

The first plant virus system used was a recombinant Tobacco Mosaic Virus (TMV) where the capsid protein was fused to a malarial epitope [53] followed by others (for review see [43, 57]). “Agro-infection” has been developed as a versatile tool for a rapid production of vaccines and antibodies in transiently expressing plant tissues, especially tobacco leaves [15, 16]. Simultaneously, a large number of expression constructs could be tested. This method can easily be scaled up by using vacuum-mediated “Agro-infiltration.” Lomonosoff and co-workers positioned a gene of interest (GOI) between the 5' leader sequence and 3' untranslated region (UTR) of RNA-2, thereby emulating a presumably stable mRNA for efficient translation. High-level expression could also be achieved in the absence of RNA-1-derived replication functions using *Agrobacterium*-mediated transient expression. Deletion of an in-frame start codon upstream of the main translation initiation site led to a massive increase in foreign protein accumulation (10–20% of total extractable protein; [47]). The magnICON<sup>®</sup> system (MagniFection) developed by Icon Genetics (Halle, Germany; now a part of Bayer Innovation GmbH, Düsseldorf, Germany) combined significant mRNA expression enhancement by a TMV-based transient expression vector with systemic delivery based on “Agro-infiltration” [24]. A recent press release announced that Bayer started clinical Phase I study with personalized vaccines from tobacco plants, produced with the magnICON<sup>®</sup> system, for treatment of non-Hodgkin's lymphoma (<http://www.icongenetics.com/html/5954.htm>). Stable transformants have been widely used to express antibodies (for review see [9]) and vaccines (for review see [52]) in transgenic plants. The expression of recombinant proteins in storage organs such as seeds [13] resulted in functional and stable products that could be stored at room temperature for extended times without significant loss in amount and activity [51]. Stable expression in

dicotyledonous seeds could be significantly boosted by specific regulatory sequences as demonstrated for scFvs [8]. Seed tissues therefore represent a very attractive target for production and extraction of pharmaceutical proteins commercially.

In addition to the expression of recombinant proteins in cultured *Nicotiana* cells (for reviews see [14, 31]), expression in duck weed (*Lemna minor*) and in moss bioreactors are alternative interesting systems providing containment during production. The duck weed system [5] and the moss bioreactors (for review see [11]) provide the possibility of glycan optimization as well. Transgenes could also be targeted to the chloroplast, ensuring that they are embedded in a chloroplast DNA homology region. The number of transgene copies after establishment of homoplasmy was shown to be very high leading to increased expression levels [7]. Epigenetic phenomena (e.g., transgene silencing) are apparently absent in chloroplasts, therefore these plant organelles offer ideal prerequisites for the production of functional vaccine antigens. Moreover, chloroplast DNA is absent in pollen, and thus limits the potential for outcrossing. Unfortunately, chloroplasts lack major post-translational modification machineries such as glycosylation (for review see [57]), and accordingly their utility is limited to molecules which do not require glycosylation.

### Edible Vaccines and Purification

The basic idea of edible vaccines was to feed animals with genetically engineered grain directly bypassing purification and complicated and expensive downstream processing.

However, this simple approach has been replaced by plant-derived vaccines because of two main reasons. Firstly, the expression level of the antigen in harvestable parts from the same plant can vary substantially. Secondly, a complete segregation of plants for pharmaceutical or veterinary applications from those meant for human or animal consumption is required [52]. However, another interesting approach making use of high-level expression of recombinant antibodies in legume seeds, e.g., peas [45], was the feeding of neutralizing recombinant antibodies against enterotoxigenic *Escherichia coli* strains to piglets. These antibodies were sufficiently active in the intestine of the fed

animals [28, 46]. A similar approach has been applied for recombinant antibodies against gastrointestinal parasites of chickens, which were expressed in peas [58].

Oral administration is not always the major route for all plant-derived vaccines. In some cases purified antigens are required for injection necessitating the development of specific purification procedures for each product. Two main challenges have to be overcome when purifying proteins from plant material: (1) impurities (proteins, carbohydrates, oils, phenolic compounds, phytic acids, nucleic acids, and other trace products) associated with each plant system must be removed, and (2) low concentrations of the target protein following initial extraction into an aqueous medium have to be avoided. Therefore, special downstream separation units are required to handle large volumes [39]. Specific methods have to be developed to achieve high amounts of vaccines and/or antibodies in a correctly folded and functional form. The successful development of such methods is also dependent on rather high and stable accumulation of the transgenic proteins *in planta*. Here, fusion to elastin-like polypeptides (ELPs) allows both easy and scalable purification as well as enhancement of the accumulation of the recombinant ELP fusion protein. Elastin-like polypeptides are highly biocompatible proteins. They exhibit the useful property of a thermally responsive reversible phase transition. These characteristics improve the efficiency with which recombinant proteins can be purified. ELP fusion proteins also exhibit reversible phase transition property. This new technology, named “ELPylation,” has recently been extended to plant cells and several plant-based expression systems have been evaluated for the production of ELPylated proteins (for review see [22]). The approach has been applied to vaccines [19], complete immunoglobulins [17, 20, 21], and antibody derivatives, scFv [49], as well as VHH [4]. For veterinary purposes, where economical features such as low price and easy-to-handle products are major factors of commercial viability, “ELPylation” is a useful component of enrichment and purification strategies.

### Future Directions

Over the past years, plant-based production of recombinant proteins has been developed and 11 plant-derived

non-pharmaceutical proteins (avidin, trypsin,  $\beta$ -glucuronidase, aprotinin, lactoferrin, lysozyme, thyroid-stimulating hormone receptor, Hantaan and Puumala viral antigens, peroxidase, laccase, and cellulase) have been brought to the market [52] indicating a huge capability of these expression technologies for the production of diagnostic and therapeutic proteins for both human and veterinary medicine. Six years after the commercialization of the first plant-derived recombinant protein, TrypZean, from corn (ProdiGene, USA), only two plant-derived compounds are in late-stage clinical trials: Interferon  $\alpha$ -2b made in aquaculture (*Lemna* expression system, LEX system) for the treatment of hepatitis C infections (Biolex Therapeutics, Pittsboro, USA) and taligurase alfa, a form of the enzyme glucocerebrosidase known as prGCD in development for treatment of Gaucher’s disease from Protalix Biotherapeutics (Carmiel, Israel). Recently, Pfizer acquired rights to prGCD produced in carrot cells and became the first big pharma company to commit itself to take to the market a biologic plant-produced drug [42].

In view of the new influenza A H1N1 pandemic, plant-based expression systems represent the fastest production for any influenza vaccine as it was demonstrated by two research groups – at Medicago Inc. (Québec, Canada) and at the Fraunhofer Institute (Plymouth, USA) – via the transient expression of the H1 HA protein in tobacco plants [43]. These results underline the advantages of plant-based expression technologies over traditional expression technologies for the production of antigens. This should be essentially true for veterinary purposes, where costs should be generally lower fitting into economical parameters of animal-based food production. Here, the “old” concept of edible vaccines could be verified much easier, because seeds could be used as a source that is an essential and common component of the feed, which do not need to be treated at harsh conditions as baking or cooking. Farm animals grown in high numbers in confined conditions are a suitable target for future attempts to produce veterinary pharmaceuticals using plant-based expression systems. Nevertheless, further improvement of expression levels and development of easy and cheap downstream processes are still needed before decisions about economic viability of transgenic plant-based pharmaceuticals for animal health could be made.

## Acknowledgments

The authors thank Stefan Rose-John and Paul Christou for a critical review of the manuscript.

## Bibliography

- Altpeter F, Baisakh N, Beachy R, Bock R, Capell T, Christou P, Daniell H, Datta K, Datta S, Dix PJ, Fauquet C, Huang N, Kohli A, Mooibroek H, Nicholson L, Nguyen TT, Nugent G, Raemakers K, Romano A, Somers DA, Stoger E, Taylor N, Visser R (2005) Particle bombardment and the genetic enhancement of crops: myths and realities. *Mol Breed* 15:305–327
- Behring E, Kitasato S (1890) Ueber das Zustandekommen der Diphtherieimmunitaet und der Tetanusimmunitaet bei Tieren. *Dtsch Med Wochenschr* 16:1113–1114
- Conrad U, Floss DM (2010) Expression of antibody fragments in transgenic plants. In: Kontermann R, Dübel S (eds) *Antibody engineering*, vol 2. Springer, Berlin, pp 377–386
- Conrad U, Plagmann I, Malchow S, Sack M, Floss DM, Kruglov AA, Nedospasov SA, Rose-John S, Scheller J (2010) ELPylated anti-human TNF therapeutic single-domain antibodies for prevention of lethal septic shock. *Plant Biotechnol J*. doi:10.1111/j.1467-7652.2010.00523.x
- Cox KM, Sterling JD, Regan JT, Gasdaska JR, Frantz KK, Peele CG, Black A, Passmore D, Moldovan-Loomis C, Srinivasan M, Cuisson S, Cardarelli PM, Dickey LF (2006) Glycan optimization of a human monoclonal antibody in the aquatic plant *Lemna minor*. *Nat Biotechnol* 24:1591–1597
- Curtiss RI, Cardineau CA (1990) Oral immunization by transgenic plants. World Patent Application, WO 90/02484
- Daniell H, Chebolu S, Kumar S, Singleton M, Falconer R (2005) Chloroplast-derived vaccine antigens and other therapeutic proteins. *Vaccine* 23:1779–1783
- De Jaeger G, Scheffer S, Jacobs A, Zambre M, Zobell O, Goossens A, Depicker A, Angenon G (2002) Boosting heterologous protein production in transgenic dicotyledonous seeds using *Phaseolus vulgaris* regulatory sequences. *Nat Biotechnol* 20:1265–1268
- De Muyneck B, Navarre C, Boutry M (2010) Production of antibodies in plants: status after twenty years. *Plant Biotechnol J* 8:529–563
- De Wilde C, Peeters K, Jacobs A, Peck I, Depicker A (2002) Expression of antibodies and Fab fragments in transgenic potato plants: a case study for bulk production in crop plants. *Mol Breed* 9:271–282
- Decker EL, Reski R (2008) Current achievements in the production of complex biopharmaceuticals with moss bioreactors. *Bioprocess Biosyst Eng* 31:3–9
- Ehrlich P (1900) On immunity, with special reference to cell life. The Croonian lecture. *Proc Roy Soc Lond* 66:424–448
- Fiedler U, Conrad U (1995) High-level production and long-term storage of engineered antibodies in transgenic tobacco seeds. *Biotechnology* 13:1090–1093
- Fischer R, Emans N, Schuster F, Hellwig S, Drossard J (1999) Towards molecular farming in the future: using plant-cell-suspension cultures as bioreactors. *Biotechnol Appl Biochem* 30:109–112
- Fischer R, Schillberg S (2004) *Molecular farming: plant-made pharmaceuticals and technical proteins*. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
- Fischer R, Vaquero C, Sack M, Drossard J, Emans N, Commandeur U (1999) Towards molecular farming in the future: transient protein expression in plants. *Biotechnol Appl Biochem* 30:113–116
- Floss DM, Conrad U (2010) Expression of complete antibodies in transgenic plants. In: Kontermann R, Dübel S (eds) *Antibody engineering*, vol 2. Springer, Berlin, pp 489–502
- Floss DM, Falkenburg D, Conrad U (2007) Production of vaccines and therapeutic antibodies for veterinary applications in transgenic plants: an overview. *Transgenic Res* 16:315–332
- Floss DM, Mockey M, Zanella G, Brosson D, Diogon M, Frutos R, Bruel T, Rodrigues V, Garzon E, Chevalyere C, Berri M, Salmon H, Conrad U, Dedieu L (2010) Expression and immunogenicity of the mycobacterial Ag85B/ESAT-6 antigens produced in transgenic plants by elastin-like peptide fusion strategy. *J Biomed Biotechnol* 2010:274346
- Floss DM, Sack M, Arcalis E, Stadlmann J, Quendler H, Rademacher T, Stoger E, Scheller J, Fischer R, Conrad U (2009) Influence of ELP fusions on the quantity and quality of a tobacco-derived HIV-neutralizing antibody. *Plant Biotechnol J* 7:899–913
- Floss DM, Sack M, Stadlmann J, Rademacher T, Scheller J, Stoger E, Fischer R, Conrad U (2008) Biochemical and functional characterization of anti-HIV antibody-ELP fusion proteins from transgenic plants. *Plant Biotechnol J* 6:379–391
- Floss DM, Schallau K, Rose-John S, Conrad U, Scheller J (2010) Elastin-like polypeptides revolutionize recombinant protein expression and their biomedical application. *Trends Biotechnol* 28:37–45
- Gahrtz M, Conrad U (2009) Immunomodulation of plant function by in vitro selected single-chain Fv intrabodies. In: Faye L, Gomord V (eds) *Methods in molecular biology: recombinant proteins from plants*. Humana Press, Totowa, pp 289–312
- Gleba Y, Klimyuk V, Marillonnet S (2005) Magniflection – a new platform for expressing recombinant vaccines in plants. *Vaccine* 23:2042–2048
- Goldstein D (1997) Banana vaccines. *Sci World* 53:4
- Gomord V, Faye L (2004) Posttranslational modification of therapeutic proteins in plants. *Curr Opin Plant Biol* 7: 171–181
- Gomord V, Fitchette AC, Menu-Bouaouiche L, Saint-Jore-Dupas C, Plasson C, Michaud D, Faye L (2010) Plant-specific glycosylation patterns in the context of therapeutic protein production. *Plant Biotechnol J* 8:564–587
- Hagemann M (2006) Untersuchungen zur Wirksamkeit oral applizierter Antikörper (produziert in transgenen Hefen und Erbsen) gegen enterotoxische *Escherichia coli* (ETEC) bei experimentell infizierten Absetzferkeln. Dissertation, Tierärztliche Hochschule Hannover

29. Hammond RW, Nemchinov LG (2009) Plant production of veterinary vaccines and therapeutics. In: Karasev AV (ed) *Plant-produced microbial vaccines*. Springer, Berlin, pp 79–102
30. Hassan S, van Dolleweerd CJ, loakeimidis F, Keshavarz-Moore E, Ma JKC (2008) Considerations for extraction of monoclonal antibodies targeted to different subcellular compartments in transgenic tobacco plants. *Plant Biotechnol J* 6:733–748
31. Hellwig S, Drossard J, Twyman RM, Fischer R (2004) Plant cell cultures for the production of recombinant proteins. *Nat Biotechnol* 22:1415–1422
32. Hiatt A, Cafferkey R, Bowdish K (1989) Production of antibodies in transgenic plants. *Nature* 342:76–78
33. Hood EE, Woodard SL, Horn ME (2002) Monoclonal antibody manufacturing in transgenic plants – myths and realities. *Curr Opin Biotechnol* 13:630–635
34. Hussack G, Grohs BM, Almquist KC, McLean MD, Ghosh R, Hall JC (2010) Purification of plant-derived antibodies through direct immobilization of affinity ligands on cellulose. *J Agric Food Chem* 58:3451–3459
35. Ko K, Brodzik R, Stepiewski Z (2009) Production of antibodies in plants: approaches and perspectives. In: Karasev AV (ed) *Plant-produced microbial vaccines*. Springer, Berlin, pp 55–78
36. Kohler G, Milstein C (1975) Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* 256:495–497
37. Mason HS, Lam DMK, Arntzen CJ (1992) Expression of hepatitis B surface antigen in transgenic plants. *Proc Natl Acad Sci USA* 89:11745–11749
38. Meeusen ENT, Walker J, Peters A, Pastoret PP, Jungersen G (2007) Current status of veterinary vaccines. *Clin Microbiol Rev* 20:489–510
39. Menkhaus TJ, Roseland J (2008) Recovery of proteins from corn and soybean extracts by membrane adsorption. *Biotechnol Prog* 24:1075–1084
40. Mor TS, Gomez-Lim MA, Palmer KE (1998) Perspective: edible vaccines – a concept coming of age. *Trends Microbiol* 6:449–453
41. Pujol M, Ramirez NI, Ayala M, Gavilondo JV, Valdes R, Rodriguez M, Brito J, Padilla S, Gomez L, Reyes B, Peral R, Perez M, Marcelo JL, Mila L, Sanchez RR, Paez R, Cremata JA, Enriquez G, Mendoza O, Ortega M, Borroto C (2005) An integral approach towards a practical application for a plant-made monoclonal antibody in vaccine purification. *Vaccine* 23:1833–1837
42. Ratner M (2010) Pfizer stakes a claim in plant cell-made biopharmaceuticals. *Nat Biotechnol* 28:107–108
43. Rybicki EP (2009) Plant-produced vaccines: promise and reality. *Drug Discov Today* 14:16–24
44. Rybicki EP (2010) Plant-made vaccines for humans and animals. *Plant Biotechnol J* 8:620–637
45. Saalbach I, Giersberg M, Conrad U (2001) High-level expression of a single-chain Fv fragment (scFv) antibody in transgenic pea seeds. *J Plant Physiol* 158:529–533
46. Saalbach I, Riehl M, Giersberg M, Kümlehn J, Falkenburg D (2007) Production of recombinant antibodies in pea seeds and their oral application in piglets. In: Xu Z (ed) *Biotechnology and sustainable agriculture 2006 and beyond*. Springer, The Netherlands, pp 399–402
47. Sainsbury F, Lomonosoff GP (2008) Extremely high-level and rapid transient protein production in plants without the use of viral replication. *Plant Physiol* 148:1212–1218
48. Santi L (2009) Plant derived veterinary vaccines. *Vet Res Commun* 33:S61–S66
49. Scheller J, Leps M, Conrad U (2006) Forcing single-chain variable fragment production in tobacco seeds by fusion to elastin-like polypeptides. *Plant Biotechnol J* 4:243–249
50. Sharma AK, Sharma MK (2009) Plants as bioreactors: recent developments and emerging opportunities. *Biotechnol Adv* 27:811–832
51. Stoger E, Ma JKC, Fischer R, Christou P (2005) Sowing the seeds of success: pharmaceutical proteins from plants. *Curr Opin Biotechnol* 16:167–173
52. Tiwari S, Verma PC, Singh PK, Tuli R (2009) Plants as bioreactors for the production of vaccine antigens. *Biotechnol Adv* 27:449–467
53. Turpen TH, Reinl SJ, Charoenvit Y, Hoffman SL, Fallarme V, Grill LK (1995) Malarial epitopes expressed on the surface of recombinant Tobacco Mosaic Virus. *Biotechnology* 13:53–57
54. Tzfira T, Citovsky V (2006) *Agrobacterium*-mediated genetic transformation of plants: biology and biotechnology. *Curr Opin Biotechnol* 17:147–154
55. Usha R, Rohll JB, Spall VE, Shanks M, Maule AJ, Johnson JE, Lomonosoff GP (1993) Expression of an animal virus antigenic site on the surface of a plant-virus particle. *Virology* 197:366–374
56. Walsh G, Jefferis R (2006) Post-translational modifications in the context of therapeutic proteins. *Nat Biotechnol* 24:1241–1252
57. Yusibov V, Rabindran S (2008) Recent progress in the development of plant-derived vaccines. *Expert Rev Vaccin* 7:1173–1183
58. Zimmermann J, Saalbach I, Jahn D, Giersberg M, Haehnel S, Wedel J, Macek J, Zoufal K, Glunder G, Falkenburg D, Kipriyanov SM (2009) Antibody expressing pea seeds as fodder for prevention of gastrointestinal parasitic infections in chickens. *BMC Biotechnol* 9:79

---

## Plant Oil Fuels Combined Heat and Power (CHP)

KLAUS THUNEKE

Technologie- und Förderzentrum im  
Kompetenzzentrum für Nachwuchsende  
Rohstoffe (TFZ), Straubing, Germany

### Article Outline

Glossary

Definition of the Subject

Introduction

Oil Processing and Purification  
 Fuel-Relevant Properties  
 Technical Fundamentals of Plant Oil-Fueled CHP  
 Plants  
 Case Study: Economic Efficiency  
 Future Directions  
 Bibliography

## Glossary

**Cogeneration** Cogeneration, also known as combined heat and power (CHP), describes the simultaneous production of both mechanical energy and useful heat from various sources of energy by a thermodynamic process in a technical plant [1, 2].

**Combined heat and power plant (CHP Plant)** A combined heat and power plant (CHP plant) or cogeneration plant provides simultaneously electricity and useful heat.

**Electricity generation efficiency** Electricity generation efficiency or electrical efficiency is the ratio between the electricity output and the energy input of an energy conversion system. Whereas small-scale CHP plants with combustion engines feature electricity generation efficiencies of roughly 30%, large-scale CHP plants obtain up to 45% [3, 4].

**Energy conversion efficiency** Energy conversion efficiency or overall conversion efficiency is the ratio between the useful energy output and the energy input of an energy conversion system. For CHP plants, energy conversion efficiency is the sum of electricity conversion efficiency and heat conversion efficiency. Electrical and thermal auxiliary power for the CHP plant (e.g., for pumps, control unit, fuel preheating, etc.) has to be deducted [3].

**Heat generation efficiency** Heat generation efficiency or thermal efficiency is the ratio between the useful heat output and the energy input of an energy conversion system. It is the result of the heat extraction by heat exchangers. Here, the temperature level is decisive [3].

**Plant oil fuel** Plant oil fuel is derived from oil-containing plant parts for the use in plant oil compatible combustion engines. In Germany, for example, plant oil fuel quality is specified by the national pre-standard DIN V 51623 (publication expected in 2011).

**Power-to-heat ratio** Power-to-heat ratio is the ratio between produced electrical and useable thermal energy. It is an evaluation criterion of CHP plants and ranges normally between 0.4 and 0.6. Because the power-to-heat ratio is increasing when heat generation efficiency is decreasing, additionally the energy conversion efficiency has to be specified [3].

**Rapeseed oil fuel** Rapeseed oil fuel is oil extracted from rapeseed, for the use in plant oil compatible combustion engines. In Germany, for example, rapeseed oil fuel quality is specified by the national standard DIN 51605 [5].

**Thermal fuel power** Thermal fuel power or rated thermal input describes the fuel heat content on basis of the net calorific value feed into an energy conversion system within a defined period of time [3].

**Utilization ratio** Utilization ratio of an energy conversion system is the ratio between the usable energy and the energy input within a given period of time.

## Definition of the Subject

Combined heat and power (CHP) or cogeneration is the simultaneous generation of both useable heat and power in a single process by a heat and power supply station or an engine. The mechanical energy is usually converted into electricity by a generator. The thermal energy can be used for heating or technical processes. CHP is a highly efficient way to use either fossil or renewable fuels and can therefore contribute significantly to reach sustainable energy goals. The benefits of CHP can be of a social, economic, and environmental nature:

- Support of local economy
- Contribution to energy security
- Saving of fossil resources
- Protection of environment and climate

Plant oils are well suited to be used as fuels for CHP with self-ignition engines, using the diesel principle. By using plant oil fuels, additional benefits can be utilized in comparison to fossil fuels. Assuming a sustainable production of the plant oil, this option can help to save fossil resources and to reduce greenhouse gas emissions. Additionally, this possibility can contribute to soil and water protection because of their high biodegradability and low ecotoxicity. Thus, plant oil fuels are

preferably to be used in environmental sensitive areas such as alpine regions or water protection areas. In rural areas with local production and use of the plant oil fuel and the coproduct press cake as fodder, a high level of closed mass flow circles can be obtained.

Plant oil-fuelled CHP plants can play an important role for a decentralized power and heat supply. In case of a favorable legal framework guaranteeing feed-in tariffs for electricity from renewable resources, economic efficiency might be given. However, an appropriate heat use concept has to be incorporated. For a reliable and low-emission operation, various aspects regarding fuel quality and technical equipment (e.g., exhaust gas after-treatment) are to be considered.

Although plant oils can be used as fuel in self-ignition engines, properties vary significantly from those of conventional diesel fuel. This applies particularly for the ignition behavior and the viscosity. To guarantee fine dispersion of the injection spray for a high combustion quality and to minimize deposit formation on injectors and pistons, a technical adaptation of CHP engines and periphery under consideration of the requirements of the plant oil fuel is essential. Measures of adaptation can include:

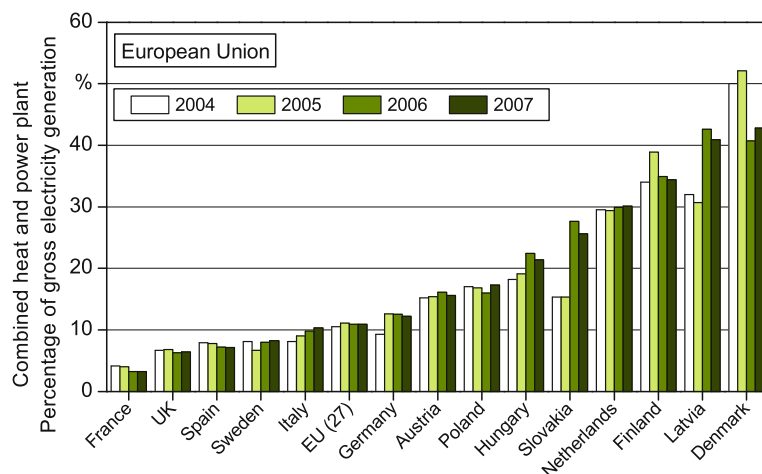
- Exchange of incompatible materials, fuel pipes, pumps, filter, injectors
- Preheating of fuel, injectors, or cooling water
- Adjustment of injection parameters

Overall technology barriers of plant oil-fuelled CHP plants are low. A high standard in operational reliability has already been achieved. Further research should aim on improvement of plant efficiency, implementation of exhaust gas after-treatment systems, and continuing standardization of promising plant oil fuels at an international level.

## Introduction

According to a Communication from the Commission of the European Union (EC) to the European Parliament and the Council, Europe can save more energy by combined heat and power generation. Cogeneration plants can contribute to energy security, sustainable energy supply, a better environment and combating climate change [6]. In addition, technology know-how opens export possibilities and offers opportunities for economic development, particularly at regional and local level. For promotion of a highly efficient combined heat and power technology based on users' heat demand a specific legal framework, the so-called Cogeneration Directive, has been introduced by the EC. Therein possible principles of support by the Member States are established.

As shown in Fig. 1, in 2007, the share of electricity from CHP of the total electricity generation in the EU-27 was 10.9%. However, the percentage varies widely between different Member States. Denmark (42.8%)



**Plant Oil Fuels Combined Heat and Power (CHP). Figure 1**

Share of CHP in total electricity generation of European Member States, EU 27 (Data source: EUROSTAT 2010 [1])



and Latvia (40.9%) feature the highest share of electricity from CHP, followed by Finland (34.4%) and the Netherlands (30.3%). Germany ranges with 12.2% in the middle, just above the European average. The share of CHP on electricity generation is lowest in France (3.2%). In Slovakia and Latvia, CHP has considerably increased between 2004 and 2007. In most other Member States, however, none or only little increase was observed.

The installed electrical power of CHP plants can range widely between some hundred megawatts at industrial scale and a few kilowatts in micro-CHP plants for private houses. Big CHP plants usually require extensive heat distribution networks for heat transport from the power station to the end user. Because of the high costs for the network and to minimize heat losses during heat distribution, compact local CHP plants installed close to the place of heat demand can be a reasonable and economic efficient alternative.

The example in Fig. 2 shows that with a fuel energy input of 100 kWh, an internal combustion engine in CHP mode can provide exemplarily about 27 kWh electrical power and 61 kWh heat. This regards to an electrical efficiency of 27%, a thermal efficiency of 61%, and an overall efficiency of 88%. However, efficiency rates vary widely between different plant designs. For example, the electrical power output can reach 40% and more if large-scale diesel engines are used. With increasing electrical efficiency, the thermal energy

output is decreasing comparatively. The losses by heat emissions, mainly through the engine system and exhaust gas, can add up to some 12 kWh.

In conventional power stations, the heat is not used in most cases. This is due to a remote large-scale production, which does normally not justify the costs of a pipe network for heat transportation and distribution to the end users. To gain the same electrical and usable thermal energy output as in the example of a CHP plant in Fig. 2, by separate heat and power generation, a fuel input of 67 kWh for the heating with burner plus 71 kWh for the power station (altogether 138 kWh) is necessary (Fig. 3). This results in primary energy savings of up to 28% by cogeneration.

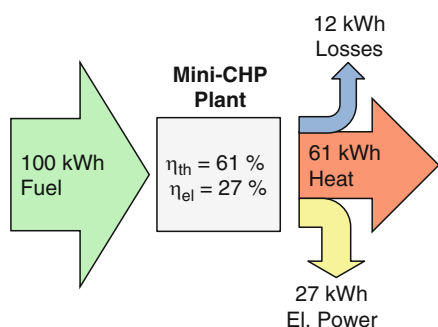
CHP plants with electrical power output of less than 1 MW are usually driven by combustion reciprocating piston engines [2]. These engines are widespread, highly developed, and can achieve up to 50,000 operating hours. Compression ignition engines for diesel fuels as drive propulsion systems for CHP plants are far more frequent than Otto-cycle engines due to their higher efficiency [3].

In diesel engines, mostly liquid fuels, such as heating oil, fatty acid methyl esters (FAME), and pure plant oils, are used. These liquid biofuels are common in Europe, even though they gain more importance outside Europe [7].

Plant oil-fuelled CHP plants are characterized by several benefits. Plant oil fuels save fossil resources and reduce greenhouse gas emissions, sustainable production provided. Because of their high biodegradability, plant oil fuels are predestinated to be used in environmental sensitive areas such as alpine regions or water protection areas. In rural areas with local production and use of the plant oil fuel (e.g., rapeseed oil) and the coproduct press cake, a high level of closed mass flow circles can be obtained. This results in high energy-utilization level and positive effects on economic regional development.

#### Case Study: Plant Oil Compatible CHP Plants in Germany

During the 1970s, engine-driven CHP plants were introduced, enabling decentralized energy supply. Since the liberalization of the energy market in the late 1990s, even private households can act as power contractors [4].



#### Plant Oil Fuels Combined Heat and Power (CHP).

Figure 2

Energy flow schema of combined heat and power (CHP) supply with small-scale CHP plants, using internal combustion engines

Today, more than 30 mainly medium-sized enterprises provide plant oil compatible CHP units in Germany. After years of increasing plant numbers with the highest growth rates between 2006 and 2007, recently the demand on plant oil compatible CHP plants is decreased strongly. According to Fig. 4, the number of plants dropped from about 2,700 in the year

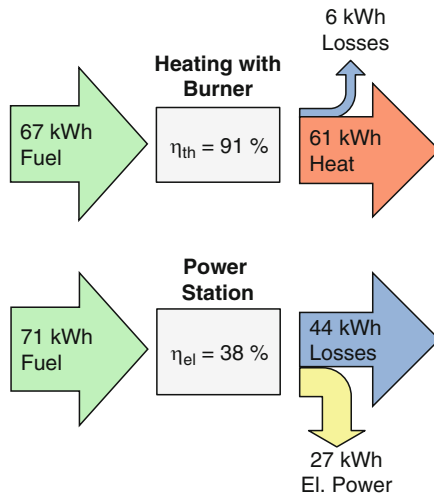
2007 to 1,400 in the year 2008, resulting in a reduction of installed electrical power from some 400 MW to 310 MW [8]. Presently, this development is continuing [9]. Suppliers of plant oil compatible CHP plants state an almost total absence of orders.

Reasons for this obvious downturn recorded for all plant sizes up to 1 MW electrical capacity were high prices for plant oil fuels in the year 2008, accompanied with inadequately performed heat use concepts of CHP units. Thus, many plants have been shut down for economic reasons. Additionally, the amendment of the Renewable Energy Act [33] led to a lack of planning reliability, regarding power feed-in tariffs as well as certification issues of sustainable plant oil fuels.

However, operators with favorable long-term contracts for plant oil fuels or CHP plants with a high degree of heat utilization often could hold up an economic viable operation. In general, a tendency from small- to large-scale CHP plants could be noted (Fig. 4).

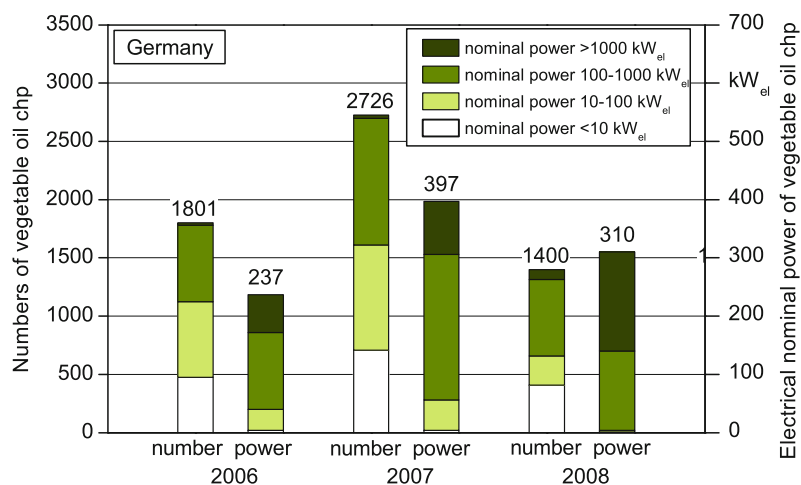
Figure 5 shows that with 88%, the highest share of the total amount of plant oil that was used in CHP plants in Germany in the year 2007 (ca. 700 Mio. l) was palm oil. With decreasing electrical nominal power, less palm oil and more rapeseed oil as well as some soy oil is used. Small-scale plant oil compatible CHP plants are predominantly fuelled with rapeseed oil.

The main reason for the leading role of palm oil in large-scale CHP plants is the lower market price of



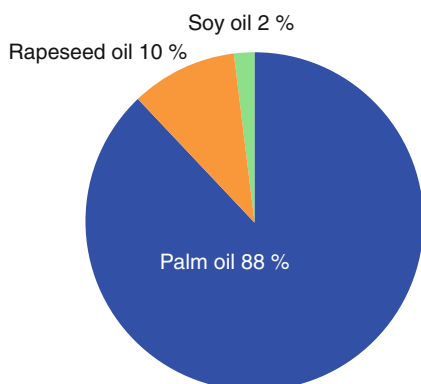
Plant Oil Fuels Combined Heat and Power (CHP). Figure 3

Energy flow schema of separate heat and power generation with burner heating and power station



Plant Oil Fuels Combined Heat and Power (CHP). Figure 4

Development of number and installed electrical power of plant oil fuelled CHP plants in Germany (Data source: DBFZ 2009 [8])



### Plant Oil Fuels Combined Heat and Power (CHP).

**Figure 5**

Share of plant oils, used as fuel for plant oil compatible CHP plants (Data source: DBFZ (2009) [8])

palm oil in comparison to rapeseed or soy oil. To ensure flowability of palm oil, extensive technical measures are necessary for heating up the entire fuel system. For small-scale CHP plants, these measures are usually too expensive in terms of installation and operation. Hence, rapeseed oil is the preferred option within the lower power range.

### Oil Processing and Purification

In Central Europe, mainly rape and sunflower oilseeds are suited to be used for the production of plant oil fuels. Storable rapeseed for instance contains about 43% oil, 40% crude protein and extractives, 7% water, 5% crude fiber, and 5% ash. The objective of the different ways of oil-processing technologies is to separate the oil content from the seed as effective as possible. Unwanted components of the seed, however, should not be transformed into the oil or need to be excluded from the oil afterwards.

For oil extraction, two different production technologies are available. Mechanical plus solvent extraction of the oilseed in industrial oil mills with processing capacities of up to 4,000 t oilseeds per day and solely mechanical extraction in local small-scale oil mills with processing capacities of 0.5–35 t per day (see [10, 11, 12, 13]).

The process steps of oil processing in industrial plants are:

- Pretreatment of the oilseed (cleaning, drying, optionally hulling, crushing, conditioning)

- Oil processing (mechanical pre-extraction, additional oil separation from extraction residues by solvents)
- After-treatment of the extraction meal (removal and recycling of solvent)
- Refining (removal of unwanted components inserted during mechanical and chemical extraction by degumming, deacidification, bleaching, and deodorization)

In small-scale oil mills, many of these process steps are omitted. Here, oil processing is characterized by following process steps:

- Pretreatment of the oilseed (cleaning, drying, optionally crushing, hulling, rolling)
- Oil extraction by cold pressing (solely mechanical oil separation, mostly by screw presses)
- Oil purification (separation of turbid substances from the oil by sedimentation, filtration, or centrifugation)
- Security filtration

Whereas oil yields in industrial large-scale oil mills obtain 99% of the seeds' oil content, only some 80% are achieved in small-scale oil mills.

Cold pressed rapeseed oil contains about 0.5–6.0% (m/m) solid substances (without oil), derived from the solid components of the oilseed. Solid substances have to be removed from the oil as completely as possible because oil purity is an important quality criteria for fuel use in combustion engines. Oil purification should be carried out at least by two purification steps: main purification and subsequent security filtration.

Because small-scale oil production does not feature any refining, especially oilseed quality, process technology as well as storage conditions have impacts on oil characteristics.

### Fuel-Relevant Properties

Rapeseed oil is the predominant plant oil for fuel use in small CHP plants. Pure rapeseed oil consists of 77–78% (m/m) Carbon (C), 11–12% hydrogen (H), and 10–11% oxygen (O). Rapeseed oil is highly biodegradable and shows only little aquatic toxicity in comparison to fossil diesel or gasoline. During storage,

reactions take place depending on storage conditions (storage tank material, temperature, oxygen, light exposure, water). Especially autoxidation and polymerization reactions are relevant. They can be minimized by suitable production and storage conditions. Under favorable conditions (darkness, cold ambient temperature of about 5°C), pure rapeseed oil can be stored at least 12 months without losing quality immoderately.

According to Table 1, characteristics of rapeseed oil according to DIN 51605 differ in important parameters from diesel fuel, heating oil, or fatty acid methyl ester (biodiesel).

In particular, the viscosity of rapeseed oil (Fig. 6), which is 10 times higher compared to fossil diesel fuel,

is often responsible for poor injection spray dispersion and insufficient combustion quality in conventional not-adapted diesel engines. As a consequence, especially during low ambient temperatures and cold starts, deposits at injectors, cylinders, pistons, or valves can occur. Similarities in rapeseed oil and diesel characteristics, however, are given among others for the net calorific value. Therefore, due to the little differences, fuel consumption (based on volume) of rapeseed oil is some 4% and in the case of biodiesel some 9% higher than that of diesel fuel or heating oil.

For a long-term reliable engine operation, the fulfillment of minimum requirements of the quality of the plant oil fuel is essential. So far, such requirements are

**Plant Oil Fuels Combined Heat and Power (CHP). Table 1** Parameters of various fuels according to DIN standards [5, 14, 15, 16, 17]

Parameter		Rapeseed oil fuel DIN 51605	Diesel fuel DIN EN 590	Heating oil EL DIN 51603-1	FAME for diesel use DIN EN 14214	FAME for heating DIN EN 14213
Density (15°C)	kg/m <sup>3</sup>	910–925	820–845	max. 860	860–900	860–900
Kinematic viscosity at 40°C at 20°C	mm <sup>2</sup> /s	max. 36.0 ca. 73.1 <sup>a</sup>	2.0–4.5	ca. 3.8 <sup>a</sup> max. 6.0	3.5–5.0	3.5–5.0
Flashpoint <sup>b</sup>	°C	min. 101	over 55	over 55	min. 101	min. 120
Sulfur content	mg/kg	max. 10	max. 50 <sup>c</sup> max. 10	>50–1,000 max. 50 <sup>d</sup>	max. 10	max. 10
Acid value	mg <sub>KOH</sub> /g	max. 2.0	n.s. <sup>e</sup>	n.s. <sup>e</sup>	max. 0.50	max. 0.50
Iodine number	g <sub>Jod</sub> /100g	max. 125	n.s. <sup>e</sup>	n.s. <sup>e</sup>	max. 120	max. 130
Oxidation stability (110°C)	h	min. 6.0	n.r. <sup>f</sup>	n.s. <sup>e</sup>	min. 8.0	min. 4.0
Ash content <sup>b</sup>	% (m/m)	n.s. <sup>e</sup>	max. 0.01	max. 0.01	max. 0.02	max. 0.02
Contamination	mg/kg	max. 24	max. 24	max. 24	max. 24	max. 24
Cetane number <sup>b</sup>	–	min. 40	min. 51	n.s. <sup>e</sup>	min. 51	n.s. <sup>e</sup>
Calorific value net gross	MJ/kg	min. 36 <sup>g</sup>	43.1 <sup>a</sup>	min. 42.6 <sup>a</sup> min. 45.4	37.1 <sup>a</sup>	min. 35

<sup>a</sup>Typical value, not specified in standard

<sup>b</sup>Different testing methods

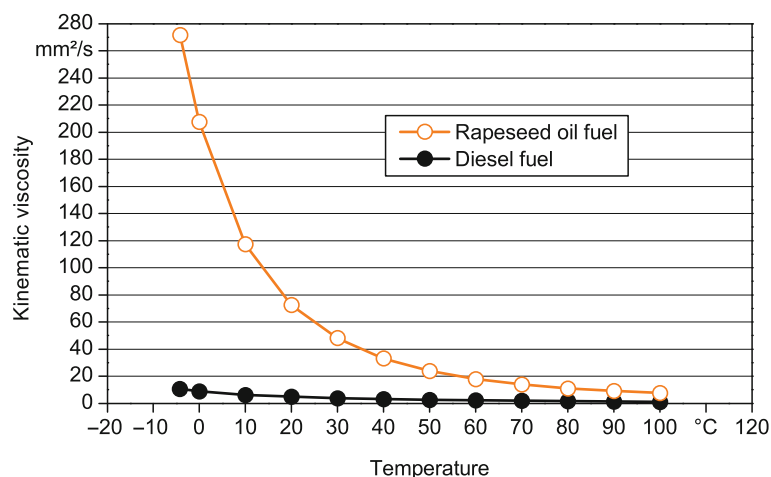
<sup>c</sup>Until 31.12.2008

<sup>d</sup>Low sulfur content heating oil

<sup>e</sup>n.s. = not specified

<sup>f</sup>n.r. = not reported due to different testing methods

<sup>g</sup>Typical value 37.5 MJ/kg



Plant Oil Fuels Combined Heat and Power (CHP). Figure 6

Kinematic viscosity of rapeseed oil and diesel fuel at different temperatures (According to Widmann et al. 1992 [18])

defined only for rapeseed oil fuel in the German standard DIN 51605. Similar to other specifications for heating oil, diesel fuel, or biodiesel, standard DIN 51605 comprises relevant product properties, test methods, and limiting values of rapeseed oil for the use as a fuel in plant oil compatible engines or heating systems.

Besides rapeseed oil, no other plant oils are investigated systematically in depth regarding their fuel-relevant properties, yet. However, many promising experiences were made in practice for palm and soy oil, as well as sunflower and jatropha oil. Table 2 shows analytical results of various plant oil samples in comparison to DIN 51605 for rapeseed oil fuel. These values derive from single analyses; no statement about their representativeness can be given.

Many requirements of DIN 51605 for rapeseed oil fuel are also fulfilled by other plant oils. Various parameters (e.g., element content) could be adjusted by the production process. However, further properties that are not relevant for rapeseed oil need to be considered individually for each type of plant oil (e.g., content of waxes for sunflower oil). This is why a simple comparison of the properties of various plant oils with limiting values defined within DIN 51605 is not sufficient for the evaluation of the suitability of a plant oil to be used as a fuel. Nevertheless, such a comparison might be of help. Palm oil with a high degree of saturation and hence high iodine number has on the one hand high

oxidation stability but, on the other hand, poor cold flow characteristics, which results in high viscosity. Thus, an external heating of all components of the fuel system is necessary, when using palm oil as a fuel.

The jatropha oil sample fulfills all limiting values of rapeseed oil fuel listed in Table 2. It could therefore be a promising alternative plant oil for fuel use. However, the specific fatty acid pattern of jatropha oil can restrain applicability. With increasing iodine number, oxidation stability is decreasing, which makes soy, sunflower, false flax, or hemp oil susceptible to oil aging. Technical problems during engine operation can be the consequence.

In order to improve fuel-relevant properties of plant oils, special oil after-treatment or addition of active agents, as it is long-term practice for fossil fuels, can be an option. Further research and standardization work, however, is necessary. Presently, a German DIN standardization committee is developing a national fuel standard for various plant oils. Publication of the pre-standard DIN V 51623 [22] is expected in 2011.

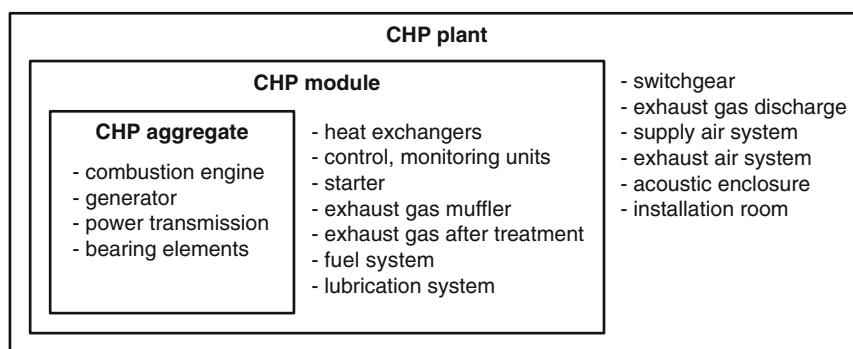
## Technical Fundamentals of Plant Oil-Fuelled CHP Plants

### Construction

CHP plants with combustion engines comprise one or more CHP modules with superior switchgear, exhaust

Plant Oil Fuels Combined Heat and Power (CHP). Table 2 Properties of various plant oil samples (not representative)

Properties	DIN 51605	Palm oil refined	Jatropha oil cold pressed	Soy oil cold pressed	Soy oil refined	Sunflower oil cold pressed	Sunflower oil refined	False flax oil cold pressed	Hemp oil cold pressed	Hemp oil refined
Reference	[5]	[19]	[20]	[19]	[19]	[19]	[19]	[19]	[21]	[21]
Density at 15°C in kg/dm <sup>3</sup>	0.910–0.925	0.93	0.92	0.92	0.92	0.92	0.92	0.92	0.93	0.93
Flash point in °C	min. 101	286	268	250	241	253	265	271	225	238
Kinematic viscosity at 40°C in mm <sup>2</sup> /s	max. 36.0	183.5	34.3	30.6	32.4	30.3	31.5	30.0	27.7	28.0
Net calorific value in MJ/kg	min. 36.0	36.7	36.8	37.0	36.9	37.0	37.0	37.1	36.5	36.7
Cetane number	min. 40	–	46	–	–	–	–	–	49	47
Iodine number in g Jod/100 g	max. 125	48	102	137	128	139	132	146	163	163
Sulfur content in mg/kg	max. 10	1.2	<1	28	<1	<1	<1	2.7	<5	<5
Acid value in mg KOH/g	max. 2.0	0.1	1.6	–	0.1	1.5	<0.1	0.3	3.9	0.2
Oxidation stability at 110°C in h	min. 6.0	27.1	10.8	4.5	7.3	3.6	5.0	3.0	2.6	2.5
Water content in mg/kg	max. 750	164	746	581	38	662	37	585	150	185



**Plant Oil Fuels Combined Heat and Power (CHP). Figure 7**

CHP-components – definition [23]

gas discharge, supply and exhaust air system, acoustic enclosure, and installation room (Fig. 7) [23].

The main component of a CHP module is the CHP aggregate consisting of combustion engine and generator with power transmission and bearing elements. Further module components are heat exchangers, control and monitoring units, starter, components of intake and exhaust gas system, fuel and lubrication system [23].

The plant oil compatible CHP plant is installed within an acoustic enclosure. The mechanical energy of the engine is converted into electrical energy by the generator. The released heat of the combustion process from the engine's cooling systems and exhaust gas is partly transferred by heat exchangers through a distribution network to the user.

Usually CHP plants are dimensioned with regard to the heat demand of a user (heat-controlled) and are operated parallel to the public electricity grid. Besides that, it is also possible to operate them power-controlled for full or partly isolated operation independent from the public electricity grid. Important for an economic efficient operation is the careful integration of the heat consumers during facility design. Insufficient heat use due to oversizing of the CHP plant can result in considerable economic losses.

### Fuel System

For medium and large-scale CHP plants, the fuel is typically stored in large tanks, which can be installed under or above ground. Depending on the local conditions, an additional intermediate storage of the fuel

can take place in smaller tanks close to the engine. Refueling of intermediate storage tanks takes place automatically by pumps and devices regulating the filling level. In small CHP plants with low fuel consumption, transportable containers with capacities of some 1,000 l often serve as fuel storage and supply.

Components of the fuel supply are pipes and flexible tubes, pressure regulators, fuel filters, fuel pumps, injection pumps, and injectors. Here, specific requirements of the plant oil on material and design has to be considered (e.g., fuel pipes and fittings made of chromated or stainless steel, flexible tubes made of flexibilizer poor or free NBR caoutchouc).

### Combustion Engine

Since the 1980s, diesel engines for the use of pure plant oil are available on the market. Plant oil compatible engines of former times designed and built especially for the use of plant oil, such as the "Elsbett engine," are no longer available on the market. During recent years, mainly series engines for diesel fuel operation have been adapted for the use of plant oil. The applied retrofitting concepts can be distinguished in single-tank and double-tank systems or single-fuel and dual-fuel systems respectively.

Double-tank systems feature a second fuel tank and fuel supply, which provides highly inflammable fuel, such as heating oil for the cold start phase. When operation temperature is reached, plant oil fuel is supplied from the main tank. Before engine turnoff, fuel from the additional tank is used again to flush the pipes and provide highly inflammable fuel in the injection

system for the following engine start. Such dual-fuel systems are often equipped with fuel preheating devices and electronically controlled fuel changers.

Single-fuel systems operate with only one type of fuel. Therefore, adaptation measures are usually more extensive for modern direct injection engines to guarantee a high spray quality and good ignition behavior of the high viscose plant oil even during cold starts and low ambient temperatures. By appropriate adaptation measures of the fuel system, combustion chamber as well as the engine management system for both direct and indirect injection series engines can be adapted with single-tank systems according to the special demands of plant oil fuel properties.

Following measures are applied in various combinations depending on engine type and adaptation concept at both single- and double-tank solutions [24, 25]:

- Exchange of components with materials not compatible for the use of pure plant oil (e.g., tubes, seals)
- Exchange of fuel pipes with little width for pipes of bigger diameter
- Exchange of fuel filters (i.e., appropriate installation of an (additional) filter)
- Exchange of fuel pump for a pump with higher power, preferably electrically driven
- Exchange or modification of injection pump, with regard to highly viscose fluids
- Exchange or modification of injectors
- Use of alternative stable materials for pistons and cylinder head
- Exchange or modification of preheating relay and heater plug (longer preheating time, placement in fuel stream, for better dispersion)
- Fuel preheating in pipes, filters, pumps, or injectors, either electrically or by warm water or oil carrying heat exchangers
- External preheating of engine by heating up engine cooling water during cold starts
- Modification of combustion chamber or valves
- Recirculation of leakage and excess fuel (possibly degasification)
- Increase of injection pressure
- Adjustment of injection time or delivery start of the pump
- Adjustment of the engine control unit

All professional adaption concepts consider adequate dimensioning and long-term stability of fuel-carrying components, such as pipes, pumps, seals, and filters. Technically sophisticated engines are advisable due to higher operational demands (higher viscosity and higher combustion temperature of plant oils). Contact with catalytic active metals, like copper, needs to be avoided, unless risking an increase of the acid value and decrease of oxidation stability of the plant oil fuel [11].

High-pressure injection systems, such as pump-injector or common rail systems, can be advantageous regarding plant oil operation, due to various possibilities of combustion process adjustment in comparison to low-pressure injection systems. Particularly, electronically regulated injection systems offer high potential of optimization.

For CHP plants, several plant oil compatible engine types are available. Small-scale CHP plants within an electrical power range up to 50 kW usually feature adapted conventional industrial diesel engines of various manufacturers. Plant oil compatible engines for CHP plants up to 500 kW electrical power are also often derived from truck or ship applications.

### Emission Reduction

For the reduction of exhaust gas emissions, various techniques are applied. These are for instance exhaust gas recirculation, oxidation catalysts, denitration catalysts, and filter systems for the removal of particulate matter.

For exhaust gas recirculation a defined branch current of exhaust gas is taken and feed into the inlet air of the engine. This leads to a decrease of the oxygen content and lower temperatures within the combustion chamber, resulting in nitrogen oxides reduction rates between 40% and 80%. With increasing recirculation rates, however, soot emissions are increasing due to less available oxygen during combustion. The oxygen content of plant oil usually allows higher recirculation rates than for diesel fuel operation.

Oxidation catalysts reduce the energy level for the induction of oxidation reactions and accelerate the reaction speed. Before the respective oxidation reaction takes place, the oxidizable substances carbon monoxide (CO) and hydrocarbons (HC) as well as oxygen are



adsorbed at the catalytic active layer. Here the molecular bonds are loosened. The optimal operation temperature for CO- and HC-conversion rates higher than 90% ranges between 200°C and 350°C. With oxidation catalysts, also aldehydes held responsible for smelly exhaust gas substances can be reduced by 80% or more. Thus, oxidation catalysts are demanded for all plant oil-fuelled CHP plants. Due to the low sulfur content of plant oils, long-term high efficient reduction rates can be secured if standard conform fuel qualities are used.

With denitration catalysts, nitrogen oxides (NO<sub>x</sub>) can be reduced effectively. For this purpose, a fluid or gaseous reduction agent (ammonia, urea or hydrocarbons) is injected into the exhaust gas stream. This so-called selective catalytic reduction (SCR) is the most common denitration technique. However, for cost reason, SCR catalysts are usually applied at engines of higher power range.

Particulate or soot filters can reduce particulate matter emissions by up to 90% and more. Also, the share of very fine particles, which are rated particularly harmful for human health, is decreased significantly. During the soot accumulation process in the filter medium, filtration surface lowers and exhaust gas counter pressure increases. Periodically, whenever the maximum tolerable pressure is reached, the particulate filter has to be regenerated by burning off the soot. The ignition of the soot can be initiated by increasing the exhaust gas temperature. Besides that, regeneration strategies can be based on continuous burning of the soot in the filter by assistance of the exhaust gas component NO<sub>2</sub>, whose share is often enlarged by upstream oxidation catalysts. For stationary engines, the

injection of burnable gas into the exhaust pipe can be another option for filter regeneration. Noncombustible ash deposits in the filter, derived from the combustion of fuel and engine lubrication oil, have to be removed from time to time by washing or air cleaning.

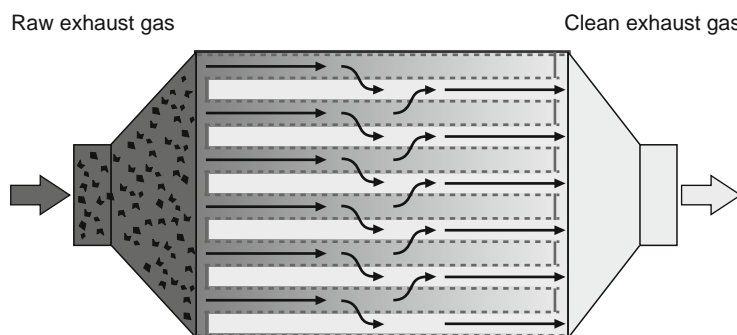
Filter systems for particulate matter emissions comprise filter medium, regeneration device, as well as a control unit [26]. As filter media high-temperature resistant materials, such as ceramic carrier substances, metals or fiber textures are used. Predominately, ceramic monoliths with alternately closed flow channels, embedded in a steel body, feature as filters (Fig. 8). Filtration takes place by enforced flow-through porous channel walls of the monolith. A catalytic layer on the filter surface can reduce soot ignition temperature considerably, enabling regeneration at lower exhaust gas temperatures.

### Exhaust Gas Pipe

The exhaust gas pipe is connected with the CHP engine over a compensator. This prevents vibration transmission and material damage by thermal shocks. The exhaust gas can be fed into an existent chimney if conforming to legal regulations. Otherwise, an isolated exhaust gas pipe of adequate length with a condensate drain has to be provided. Pipes should be of stainless steel to avoid corrosion.

### Generator and Electrical Connection

The mechanical energy of the engine is transferred into electrical energy by a generator. To operate a CHP plant independently from the electricity network, for example,



Plant Oil Fuels Combined Heat and Power (CHP). Figure 8  
 Schema of a particulate filter with alternately closed flow channels

for emergency power supply, a synchronous generator is mainly used. A synchronization device ensures that voltage, frequency, and phase of the generator and grid correspond with each other before connection.

In comparison to synchronous generators, asynchronous generators are often more robust, require less maintenance, and are cheaper in the lower power range. Because asynchronous generators require inductive idle power from the electrical network, they are usually not applicable for stand-alone operation.

The electrical connection can normally take place at the low-voltage power grid if the electrical power of the plant is lower than about 1 MW. Otherwise, a medium-voltage power grid is used.

### Heat Exchanger

Besides heat exchangers, which are flown through discontinuously (regenerators), for CHP plants mainly heat exchangers are installed, which are continuously flown through, so-called recuperators. Recuperators are distinguished as counter-flow, co-flow, or cross-flow heat exchangers, depending on the principle of operation. According to demand, they are designed as pipe bundle, plate, bag, or spiral pipe heat exchangers. The dimensioning is done in terms of necessary temperatures of the supply and return flow [3].

CHP plants can feature several heat exchangers in series for transfer of heat from the charge air, the generator cooling water, the engine cooling water, and the exhaust gas into the heating water. Furthermore, heat extraction can also take place in two separate heating circuits for realizing a higher temperature level in one of the circuits.

About 25–30% of the energy contained in the fuel is discharged via the cooling water, and an additional 30% is released as heat of exhaust gas. Heat release via the hot engine surfaces, the oil pan, or exhaust gas pipes is hardly usable. Thus, it is tried to minimize it [3].

Operational deposits in heat exchangers require periodically cleaning of the heat exchange surfaces for an effective heat transmission. Some exhaust gas heat exchangers have to be cleaned with soot brushes regularly. Other types are discontinuously or continuously self-cleaning. Maintenance on exhaust gas heat exchangers can be significantly reduced by installation of exhaust particle filter systems.

### Peak Load Boilers

Small size combined heat and power plants are predominantly operated heat-controlled, responding to the basic heat demand. Additionally, installed peak load boilers are switched on when heat delivery of the CHP plant is not sufficient to attain requested temperature in the heating system. To compensate heat peak loads during the day, a heat accumulator is established. Out of this so-called buffer, the heat demand of the users is supplied.

Heat accumulator design depends on the thermal power of the CHP plant, on the heat demand, and the usable temperature difference. The latter is limited by the maximum allowed temperature of the return flow back to the CHP plant. To secure sufficient cooling of the engine, the temperature of the return flow need to range at a maximum of 60–70°C.

### Planning and Operation

For planning and design of plant oil compatible CHP plants, generally the same principles apply as for aggregates fuelled with diesel or heating oil. These principles are described for example in the German guideline VDI-Richtlinie 3985 “Grundsätze für Planung, Ausführung und Abnahme von Kraft-Wärme-Kopplungsanlagen mit Verbrennungskraftmaschinen” [23]. However, for economic calculation, different assumptions have to be made in comparison to heating oil aggregates, concerning investment and fuel costs as well as achievable revenues for power feed-in or saving of electricity procurement costs.

The following tasks for planning CHP plants can be specified:

- Conduction of a pilot survey, demand analysis, and inventory
- Preparation of a CHP concept (module selection, operating mode)
- Check of economic efficiency of concepts
- Preliminary planning, schematic design on basis of a preliminary decision
- Interview with approving authority
- Implementation planning and preparation of tender documents and technical specifications
- Obtaining of a technical immission control report, if necessary

## Emissions

Due to high utilization of primary energy, less CO<sub>2</sub> is emitted by cogeneration plants than by a separate supply of power and heat. When using plant oil as a fuel, an additional CO<sub>2</sub> reduction is achieved. The emission of further exhaust gas components, like CO, NO<sub>x</sub>, HC, and particulate matter, however, has to be considered individually [3].

By combusting plant oil fuels, a different emission behavior is expected compared to conventional fossil fuels. Besides fuel properties, also operation mode, engine and exhaust gas after-treatment, as well as specific fuel related plant design have major influences.

Professional engine adaptation usually results in lower emissions of CO, HC, particulate matter, and PAH when using rapeseed oil fuel. NO<sub>x</sub> and aldehydes concentrations are usually little higher compared to diesel or heating oil. During low load operation or by using incompatible engines, also a converse effect can be observed [27, 28, 29, 30, 31, 32].

## Case Study: Economic Efficiency

Economic efficiency of CHP plants mainly depends on the achievable prices or credits for generated electricity and heat. The reimbursement for electricity from plant oil-fuelled CHP plants is regulated by law in several countries (e.g., Germany). The rate of reimbursement depends, e.g., on the installed electrical nominal power of the plant, the year of commissioning, the fuel, and the heat use.

All expenses of the CHP plant are usually referred to the heat. The specific heat generation costs result from the total costs per year, less the compensation for the generated energy divided by the produced usable heat quantity per year.

In Table 3, heat generation costs are calculated exemplarily for three scenarios: “CHP 1,” “CHP 2,” and “CHP 3” (without peak load boiler, buffer vessel, and planning). The three scenarios are characterized by three different plant sizes (8, 20 and 50 kW<sub>el</sub>). According to Table 3, the heat production costs range from 8.5 Cent/kWh (CHP 3) to 14.1 Cent/kWh (CHP 1).

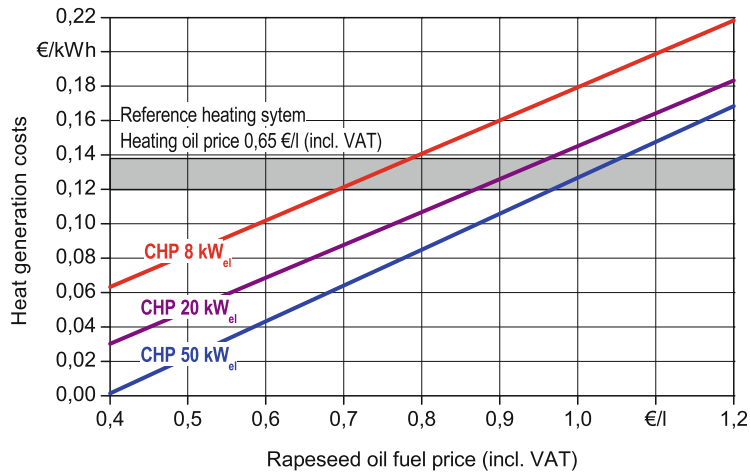
**Plant Oil Fuels Combined Heat and Power (CHP). Table 3** Economic efficiency of plant oil-fuelled CHP plants – model calculation

Assumptions		CHP 1	CHP 2	CHP 3
<i>Assumptions:</i>				
CHP nominal electrical power	kW <sub>el</sub>	8	20	50
CHP nominal thermal power	kW <sub>th</sub>	16	35	67
Investment for CHP module <sup>a</sup>	€	20 000	32 000	55 000
Investment for structure (tank, exhaust pipe) <sup>a,b</sup>	€	10 400	26 000	65 000
Costs of maintenance per year for CHP module	% of Invest.	9.0	9.0	9.0
Costs of maintenance per year for structure	% of Invest.	1.5	1.5	1.5
Labor/administrative costs per year <sup>c</sup>	% of Invest.	2.5	2.5	2.5
Insurance costs per year <sup>d</sup>	% of Invest.	1.5	1.5	1.5
Fuel consumption at nominal power	l/h	3.1	6.7	14.0
Fuel costs	€/l	0.80	0.80	0.80
Operating hours at nominal power per year	h	4 000	4 000	4 000
Power feed-in credit (0.2046 €/kWh <sub>el</sub> ) [33]	€/a	6 547	16 368	40 920
Heat generation costs (incl. power feed-in credit)	€/kWh	0.141	0.107	0.085

<sup>a</sup>Interest rate: 6%, assumed useful life 15 years for CHP module and 25 years for structure

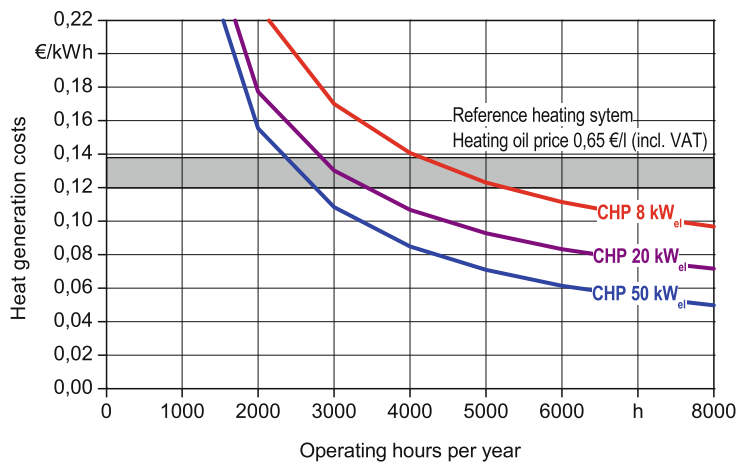
<sup>b</sup>Investment for structure: 1,300 €/kW<sub>el</sub>

<sup>c</sup>Related to investment for CHP module



Plant Oil Fuels Combined Heat and Power (CHP). Figure 9

Economic efficiency of plant oil fuelled CHP plants – Influence of the fuel price



Plant Oil Fuels Combined Heat and Power (CHP). Figure 10

Economic efficiency of plant oil-fuelled CHP plants – Influence of the operating hours

Figure 9 shows heat generation costs for the three scenarios in dependence of the fuel costs. Under the assumptions of the calculation model (Table 3), an increase of rapeseed oil fuel costs of 10 Cent/l results in an increase of the heat generation costs of around 2 Cent/l. The break-even point, where economic efficiency is solely reached by power feed-in payment and heat is virtually available for “free,” lies for CHP 3 with 50 kW<sub>el</sub> as the best option at a fuel price of 40 Cent/kWh. This shows, as expected, that for an economic efficient operation of plant oil compatible CHP plants, heat needs to be of monetary value.

A comparison of the heat generation costs of rapeseed oil-fuelled CHP plants with a reference heating system of the same power range (heating oil price of 65 Cent/l) shows some advantages for CHP plants of a minimum power of 20 kW<sub>el</sub>. However, CHP plants usually require peak load boilers or buffer vessels for short-time heat storage, which are not included in the calculation. Furthermore, these advantages only apply, when a high number of operating hours per year can be achieved.

Figure 10 shows the influence of the yearly operating hours on the heat generation costs. As it can be

seen, depending on the size of the CHP plant, only at about 3,000–5,000 operating hours per year, heat generation costs of the reference heating system (with fossil heating oil) are obtained. With decreasing operating hours, heat generation costs increase disproportionately. Thus, an economic efficient operation of rapeseed oil-fuelled CHP plants is only possible if the heat use concept is carefully considered.

### Future Directions

Climate protection and security of supply with energy sources at reasonable prices is one of the major challenges today. One approach is the extension of the local and regional decentralized energy supply. Cogeneration plants for local heat production contribute to the saving of declining resources and greenhouse gas emissions due to efficient conversion technology with overall conversion efficiencies of up to 90%.

Plant oil-fuelled CHP plants feature also additional environmental benefits. The use of plant oils contribute to soil and water protection. Because of their high biodegradability and low ecotoxicity, plant oils are predestinated to be used in environmental sensitive areas, such as alpine regions or water protection areas.

Market relevance of plant oil-fuelled CHP plants is depending on respective framework conditions. Barriers are economic insecurities due to volatile fuel prices, frequent changes of regulations, and incentive programs. Furthermore, administrative and cost expenses increase, which affect particularly small-scale CHP plants. Besides that, biofuels are still subject to oppositional discussions about social and environmental impacts. Objective debates on the perspectives of all sustainable biofuels with regard to their optimized utilization paths have to be continued. A premature commitment to certain energy sources hinders the development of a stable market for renewable energy supply.

Nevertheless, there is already a broad consensus in the need of implementing measures to reduce GHG-emissions and energy dependency. Plant oil-fuelled CHP plants can make a contribution to meet the renewable energy targets. The high potential of pure plant oils can be utilized in the short term due to existing technology. Sustainably produced plant oil fuels for cogeneration are available.

The technology barriers of plant oil-fuelled CHP are little, and a high standard in operational reliability is already achieved. Depending on the existing frame conditions, economic efficiency is given when the heat can be used reasonably during a long period of the year. Further research should aim on improvement of plant efficiency, implementation of exhaust gas after-treatment systems, and continuation of standardization of promising plant oil fuels.

### Bibliography

#### Primary Literature

1. Eurostat (2010) <http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home>. Accessed 13 April 2010
2. Schaumann G, Schmitz KW (2010) Kraft-Wärme-Kopplung, 4th edn. Springer, Auflage/Berlin/Heidelberg
3. Klausmann H (2000) Aufbau und Einsatz von anschlussfertigen BHKW-Kompaktmodulen bis 250 kW. Vulkan, Essen, p 344
4. Suttor W, Johler M, Weisenberger D (2009) Das Mini-Blockheizkraftwerk. Eine Heizung, die auch Strom erzeugt., vol 5, Auflage. C. F. Müller, Heidelberg/München/Landsberg/Frechen/Hamburg, p 140
5. Deutsches Institut für Normung (2010) DIN 51605: Kraftstoffe für pflanzenölaugliche Motoren – Rapsölkraftstoff – Anforderungen und Prüfverfahren. Beuth, Berlin
6. Europäische Kommission (2008) "Mitteilung der Kommission an das Europäische Parlament und den Rat – Mehr Energie einsparen in Europa durch Kraft-Wärme-Kopplung" /\*KOM/2008/0771 endg.\*/ vom 13.11.2008, Brüssel
7. Thrän D, Müller-Langer F (2008) Bewertung verschiedener flüssiger biogener Kraftstoffe für die stationäre Anwendung – Potenziale, Ökonomie, Ökologie. In: Blockheizkraftwerke 2008. Im Focus biogener Brennstoffe – Technik - Betriebserfahrungen, VDI-Berichte 2046. VDI Verlag, Düsseldorf, p 220
8. Deutsches BiomasseForschungszentrum (DBFZ) (2009): Monitoring zur Wirkung des Erneuerbare-Energien-Gesetzes (EEG) auf die Entwicklung der Stromerzeugung aus Biomasse – Zwischenbericht "Entwicklung der Stromerzeugung aus Biomasse 2008" im Auftrag des BMU, Eigenverlag, p 59
9. Deutsches BiomasseForschungszentrum (DBFZ) (2010) Monitoring zur Wirkung des Erneuerbare-Energien-Gesetz (EEG) auf die Entwicklung der Stromerzeugung aus Biomasse – Zwischenbericht im Auftrag des BMU, Eigenverlag, p 91
10. Remmele E (2009) Handbuch. Herstellung von Rapsölkraftstoff. In: Dezentralen Ölgewinnungsanlagen. nachwachsende-rohstoffe.de, 2. Aufl. Gülzow: Fachagentur Nachwachsende Rohstoffe e.V., p 88
11. Widmann B, Kaltschmitt M, Münch EW, Müller-Langer F, Remmele E (2009) Produktion und Nutzung von Pflanzenölkraftstoffen. In: Kaltschmitt M, Hartmann H, Hofbauer H (eds) Energie aus Biomasse – Grundlagen, Techniken und Verfahren. 2. Aufl. Springer, Berlin, pp 711–736

12. Brenndörfer M, Graf T (2005) Anlagentechnik der Ölabbpressung. In: Kuratorium für Technik und Bauwesen in der Landwirtschaft e. V. (KTBL) (Hrsg.): Dezentrale Ölsaatenverarbeitung, Nr. 427. Landwirtschaftsverlag, Münster, pp 31–36
13. Bockisch M (1993) Nahrungsfette und -öle. Eugen Ulmer, Stuttgart, p 694
14. Deutsches Institut für Normung (2009) DIN EN 590: Kraftstoffe für Kraftfahrzeuge – Dieseldieselmotoren – Anforderungen und Prüfverfahren. Beuth, Berlin
15. Deutsches Institut für Normung (2003) DIN EN 14213: Heizöle – Fettsäure-Methylester (FAME) – Anforderungen und Prüfverfahren. Beuth, Berlin
16. Deutsches Institut für Normung (2010) DIN EN 14214: Kraftstoffe für Kraftfahrzeuge – Fettsäure-Methylester (FAME) für Dieselmotoren – Anforderungen und Prüfverfahren. Beuth, Berlin
17. Deutsches Institut für Normung (2008) DIN 51603-1: Flüssige Brennstoffe – Heizöle – Teil 1: Heizöl EL, Mindestanforderungen. Beuth, Berlin
18. Widmann BA, Apfelbeck R, Gessner BH, Pontius P (1992) Verwendung von Rapsöl zu Motortreibstoff und als Heizölersatz in technischer und umweltbezogener Hinsicht. "Gelbes Heft" Nr. 40. Bayerisches Staatsministerium für Ernährung, Landwirtschaft und Forsten, München, p 650
19. Meierhofer T (2006) Untersuchungen zur Eignung verschiedener Pflanzenöle als Kraftstoff in pflanzenölauglichen BHKW. Diplomarbeit. Fachhochschule Amberg-Weiden, Fachbereich Maschinenbau/Umwelttechnik. Eigenverlag, p 157
20. ASG Analytik-Service Gesellschaft mbH (2007) Prüfbericht (unpublished)
21. Emberger P, Thuncke K, Haas R, Remmele E (2007) Prüfung von Hanföl hinsichtlich seiner Eignung als Kraftstoff für pflanzenölaugliche Motoren. Straubing: Technologie- und Förderzentrum im Kompetenzzentrum für Nachwachsende Rohstoffe, 57 Seiten, [www.nova-institut.de](http://www.nova-institut.de). Accessed 13 April 2010
22. Deutsches Institut für Normung (2010) DIN V 51623: Kraftstoffe für pflanzenölaugliche Motoren – Pflanzenölkraftstoff – Anforderungen und Prüfverfahren. Beuth (in preparation), Berlin
23. Verein Deutscher Ingenieure (2004) VDI-Richtlinie 3985: Grundsätze für Planung, Ausführung und Abnahme von Kraft-Wärme-Kopplungsanlagen mit Verbrennungskraftmaschinen. VDI-Gesellschaft Energietechnik (Hrsg.). Beuth, Berlin
24. Remmele E (2002) Standardisierung von Rapsöl als Kraftstoff – Untersuchungen zu Kenngrößen, Prüfverfahren und Grenzwerten. Dissertation an der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München, Lehrstuhl für Landtechnik, Freising-Weihenstephan, VDI-MEG Schrift Nr. 400, p 194
25. Thuncke K, Remmele E, Widmann B (2002) Pflanzenölbetriebene Blockheizkraftwerke – Leitfaden. Materialien Umwelt & Entwicklung Bayern 170. Bayerisches Staatsministerium für Landesentwicklung und Umweltfragen, München, p 66
26. van Basshuysen R, Schäfer F (2004) Lexikon Motorentchnik. Friedr. Vieweg & Sohn Verlag/GWV Fachverlage GmbH, Wiesbaden, p 1078
27. Thuncke K (2007) Verbrennung und Emissionen von Pflanzenölen in Dieselmotoren – Übersicht zu ausgewählten aktuellen Forschungsergebnissen. In: Nova-Institut GmbH (Hrsg.): Erster Internationaler Kongress zu Pflanzenölkraftstoffen, Tagungsband / Proceedings, Messe Erfurt, 6–7 September 2007. Nova Institut, Hürth, pp 189–204
28. Czerwinski J, Zimmerli Y, Neubert T, Kasper M, Mosimann M (2006) Analysis of (Nano) Particles with GTL, RME & ROR on a Modern HD-diesel engine. Report for Swiss Petrol Union, Zürich, Schweiz. Berner Fachhochschule, Hochschule für Technik und Informatik HTI, Abgasprüfstelle, Nidau (Hrsg.), Eigenverlag, p 40
29. Dobiasch A (2000) Einfluss der chemischen und physikalischen Eigenschaften von regenerativen Kraftstoffen auf das Emissionsverhalten von Verbrennungsmotoren. Fortschritt-Berichte VDI, Reihe 12, Band 428. VDI Verlag GmbH, Düsseldorf, p 173
30. Meyer M (2007) Rapsölkraftstoff in der Schweiz – Statusbericht zur Motorenforschung. In: Tagungsband zum 16. Symposium Bioenergie – Festbrennstoffe, Flüssigkraftstoffe, Biogas. Ostbayerisches Technologie-Transfer-Institut e. V. (OTTI) (Hrsg.), OTTI e. V., Eigenverlag, Regensburg
31. Prankl H, Krammer K, Janetschek H, Roitmeier T (2005) Blockheizkraftwerke auf Pflanzenölbasis. Abschlussbericht zum Forschungsprojekt BLT 012951. Eigenverlag, Wieselburg, p 106
32. Thuncke K (2009) Untersuchungen zu Abgasemissionen und zum Einsatz von Partikelfiltersystemen bei rapsölbetriebenen Blockheizkraftwerken. Dissertation an der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München, Lehrstuhl für Agrarsystemtechnik. Freising-Weihenstephan. VDI-MEG Schrift Nr. 478, p 191
33. Erneuerbare-Energien-Gesetz vom 25. Oktober 2008 (BGBl. I S. 2074), zuletzt geändert durch das Gesetz vom 11. August 2010 (BGBl. I S. 1170)

## Books and Reviews

- BINE Informationsdienst FIZ Karlsruhe (2009) Blockheizkraftwerke. Ein Leitfaden für den Anwender., 7th edn. Solarpraxis AG, Berlin
- Kaltschmitt M, Hartmann H, Hofbauer H (2009) Energie aus Biomasse. Grundlagen, Techniken und Verfahren., vol 2, Auflage. Springer, Berlin/Heidelberg
- Klausmann H (2000) Aufbau und Einsatz von anschlussfertigen BHKW-Kompaktmodulen bis 250 kW. Vulkan, Essen
- Mollenhauer KH, Tschöke H (2010) Handbook of diesel engines. Handbuch – Dieselmotoren. 2. Auflage. Springer, Berlin, p 1069
- Otto F, Otte J, Raatz A (2009) Untersuchung der Marktchancen, Hemmnisse und Systemoptionen für Strom erzeugende Heizungen vor dem Hintergrund neuer nationaler und internationaler technischer Entwicklungen im Bereich der Kleinst-BHKW im Hinblick auf zukünftig anstehende Neu- und Umstrukturierungen der deutschen Stromversorgung – Abschlussbericht. Fraunhofer IRB Verlag, Stuttgart

- Pauleickhoff W (2009) Optimierung der Wirtschaftlichkeit von Miniblockheizkraftwerken. Diplomica Verlag GmbH, Hamburg
- Remmele E (2009) Handbuch. Herstellung von Rapsölkraftstoff. In: dezentralen Ölgewinnungsanlagen. 2. Auflage. Fachagentur Nachwachsende Rohstoffe e.V., Gülzow
- Suttor W, Johler M, Weisenberger D (2009) Das Mini-Blockheizkraftwerk. Eine Heizung, die auch Strom erzeugt. 5. Auflage. Müller, Heidelberg/München/Landsberg/Frechen/Hamburg
- Thomas B (2007) Mini-Blockheizkraftwerke. Grundlagen, Gerätetechnik, Betriebsdaten., vol 1, Auflage. Vogel Buchverlag, Würzburg
- VDI-Gesellschaft Energietechnik (2008) Blockheizkraftwerke 2008. Im Focus biogener Brennstoffe. Technik – Betriebserfahrungen. VDI-Berichte 2046. VDI Verlag, Düsseldorf

## Plasma-Assisted Waste-to-Energy Processes

NICKOLAS J. THEMELIS<sup>1</sup>, ARMELLE M. VARDELLE<sup>2</sup>

<sup>1</sup>Earth Engineering Center, Columbia University, New York, NY, USA

<sup>2</sup>Laboratoire Sciences des Procédés Céramiques et de Traitements de Surface, University of Limoges, Limoges Cedex, France

### Article Outline

Glossary  
 Definition of the Subject and Its Importance  
 Introduction  
 Composition and Chemical Heat Content of MSW  
 Thermal Plasma Torches  
 Energy and Material Balances in Plasma-Assisted Gasification of MSW  
 Plasma-Assisted Processes for Treating MSW  
 Environmental Impacts  
 Future Directions  
 Bibliography

### Glossary

**Arc plasma** A gas that is heated electrically to temperatures up to 20,000 K by means of an arc struck between two electrodes.

**Arc plasma torch** Device used to generate a thermal plasma.

**Efficiency of energy generation** Ratio of net electrical energy generated to chemical heat input, per ton of MSW processed.

**MSW** Municipal solid waste, mixed waste that is collected by a given collection system.

**Non-transferred arc plasma torch** The two electrodes located within a water-cooled plasma torch.

**Torch thermal efficiency** Ratio of enthalpy input to the plasma-forming gas to electrical energy input to the plasma torch.

**Transferred arc** The material to be processed serves as an electrode.

**Vitrification** Also called glassification: converting WTE ash to a glassy substance by melting at high temperatures.

**WTE** Acronym for waste-to-energy, i.e., thermal treatment of solid wastes to recover their chemical energy content.

### Definition of the Subject and Its Importance

The thermal plasma technology [1, 2] has been used for over 30 years mainly for surface coating, metal welding and cutting, powder treatment and synthesis, and metal melting and smelting. More recently, thermal plasmas have also been used for the pyrolysis of hazardous liquids and gasses and the compaction of solid wastes [3]. Examples of the latter technology are the destruction of asbestos-contaminated waste materials and the vitrification of the ash by-product of waste-to-energy plants. Efforts to apply plasma in the thermal treatment of municipal solid wastes (MSW), in the absence of partial combustion, have not been successful because of the required high “investment” of electricity per unit of mass treated. Therefore, in this essay, we are examining processes where thermal plasma is used *in conjunction* with partial oxidation and gasification of the organic compounds contained in the MSW, thus reducing the consumption of electricity per ton of material treated. Such processes exist and are in different stages of development. Collectively, they can be called “plasma-assisted WTE” technologies and are the subject of this essay.

### Introduction

Although both conventional waste-to-energy (WTE) and plasma-assisted WTE processes involve a certain degree of combustion, there is an important

difference: in conventional WTE, carbonaceous materials are combusted and the heat of combustion is transferred to steam that powers a turbine generator of electricity. Per ton of MSW processed, the ratio of net electric energy generated for the grid/input chemical energy contained in MSW is called the thermal efficiency of energy generation.

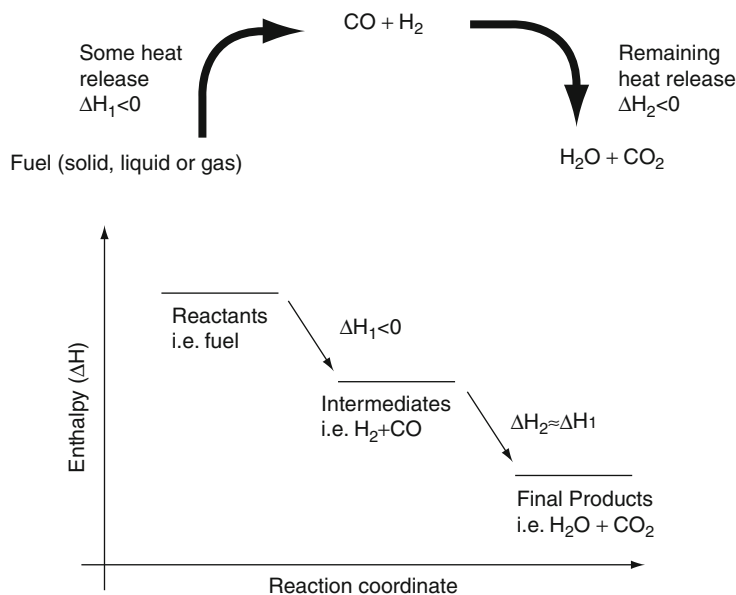
In the case of plasma-assisted gasification, the objective is to partially oxidize the carbon content of MSW to carbon monoxide (CO) and produce a synthetic gas that contains as high as possible concentrations of carbon monoxide and hydrogen and, conversely, a very low concentration of carbon dioxide. This synthetic gas or “syngas” can generate electricity in a gas engine or turbine at a higher thermal efficiency than a steam turbine. The heat generated by the plasma torch is used to provide some of the heat for gasification, to break down long hydrocarbons to CO and H<sub>2</sub>, and to vitrify the ash product of the gasification process. The syngas product can either be quenched by means of a water stream, or passed through a heat recovery exchanger to produce steam used in a steam turbine to produce additional electricity.

The main difference between traditional combustion and gasification is the amount of oxygen used. Full

combustion of the carbonaceous materials in MSW requires a large excess of air, typically 70–100% of the stoichiometric requirement. Gasification is carried out with a sub-stoichiometric amount of oxygen in order to produce mostly carbon monoxide and hydrogen. Of course, in order to utilize the energy content of the syngas, it is necessary to combust it in a gas engine or turbine. The potential advantages of gasification are that a much lower excess oxygen is required, thus simplifying the gas control system; also, the thermal efficiency of the gas engine can be substantially higher than that of the steam turbine used in grate combustion WTE. Figure 1 shows the reaction sequence in gasification systems [4].

### Composition and Chemical Heat Content of MSW

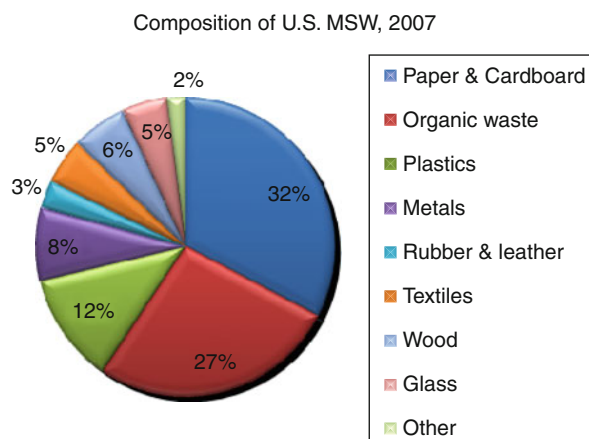
The calorific value of MSW varies considerably, depending on their content of food and green wastes, that introduce moisture, and petrochemical wastes, such as plastics and textiles, that are associated with higher calorific values than paper fiber and wood. Metals, glass, and other inorganic materials in the MSW do not contribute to its heating value; in fact they reduce it somewhat. Figure 2 shows the typical composition of U.S. MSW in 2007 [4].



Plasma-Assisted Waste-to-Energy Processes. Figure 1

Conceptual pathway for conversion of carbon fuels to final gaseous products





**Plasma-Assisted Waste-to-Energy Processes. Figure 2**  
Composition of the U.S. MSW [4]

On the basis of the MSW composition (e.g., Fig. 2) and the chemical composition of the constituent materials, it has been shown [5] that the chemical structure of *organic* compounds (both biogenic and fossil-based) in MSW can be approximated by the formula  $C_6H_{10}O_4$ ; it is interesting to note that there are about ten organic compounds in nature that have the same chemical structure [6]. In the absence of moisture and inorganic materials,  $C_6H_{10}O_4$  would have a calorific value of about 18,500 MJ per kilogram. However, since mixed MSW contains other non-combustible materials, the resulting calorific value can be expressed as follows [5]:

Heating value of mixed MSW

$$\begin{aligned}
 &= (\text{heating value of combustibles}) \times X_{\text{comb}} \\
 &\quad - (\text{heat loss due to water in feed}) \times X_{\text{water}} \\
 &\quad - (\text{heat loss due to glass in feed}) \times X_{\text{glass}} \\
 &\quad - (\text{heat loss due to metal in feed}) \times X_{\text{metal}}
 \end{aligned}$$

where  $X_{\text{comb}}$ ,  $X_{\text{water}}$ , etc are the mass fractions of combustible matter, water, etc in the MSW and  $X_{\text{comb}} + X_{\text{water}} + X_{\text{glass}} + X_{\text{metal}} = 1$ .

Substituting numerical values for the heat of reaction, evaporation of moisture, and heat carry over in the WTE ash results in the following equation.

$$\begin{aligned}
 \text{Heating value of MSW} &= 18.5 \cdot X_{\text{comb}} - 2.6 \cdot X_{\text{water}} - 0.6 \cdot X_{\text{glass}} \\
 &\quad - 0.5 \cdot X_{\text{metal}} \text{ (MJ/kg)}
 \end{aligned}$$

The effect of moisture on the calorific value of several types of solid wastes is shown graphically in Fig. 3.

In the case of conventional WTE processes, the inlet moisture in the MSW is evaporated, the organic compounds in MSW are completely combusted, and the inorganic compounds end up in the WTE ash. The exothermic chemical reaction of U.S. MSW in the combustion chamber can be represented by the following equation:

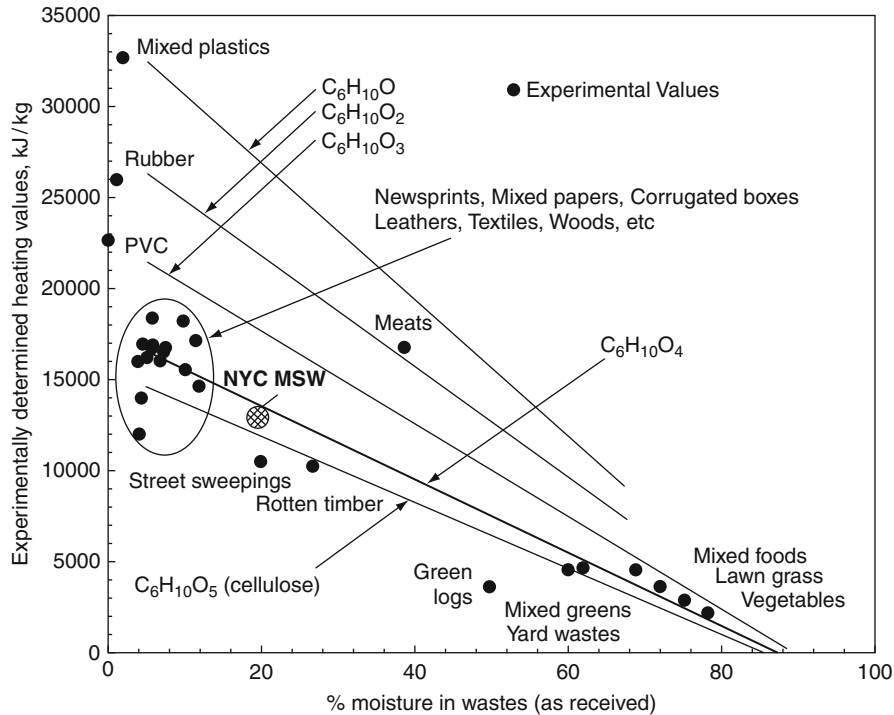


A study of 97 incinerator power plants in the European Union (Fig. 4, [7]) showed that the average MSW combusted in the E.U. has an average calorific value of 10 MJ/kg (about 2,800 kWh/t of feed). Therefore, a WTE grate combustion facility operating at a net thermal efficiency of 22% would provide 600 kWh/t to the grid. For WTE plants that provide electricity and district heating, as is done in northern European countries, the thermal efficiency can be significantly higher.

### Thermal Plasma Torches

The most common way to create a thermal plasma jet is by heating a gas stream to high temperatures (up to 20,000 K) by means of a DC- or AC-sustained electric arc between the cathode and the anode of the torch. The plasma jet is a mixture of ions, electrons and neutral particles and, because of its high temperature, can vaporize and destroy any chemical compound if the material is properly "mixed" with the gas phase. The partly ionized gas exits the plasma torch at high velocity, thus creating what is called a plasma jet. The main advantages of thermal plasma are high energy density, resulting in extremely high heat and mass transfer rates, compact size of the heat source, and rapid start-up and shut down. However, the use of electricity is a drawback since it is an expensive form of energy. Furthermore, the plasma torch needs to be water-cooled, thus introducing a heat loss that ranges from 10% to 40% depending on the torch configuration.

Thermal plasmas can also be generated by radio frequency induction and microwaves. However, for the treatment of waste, plasma is preferentially generated by DC electric discharge with two kinds of torch configurations, non-transferred and transferred arc. Combining classic gasification with plasma technology allows a higher efficiency in the production of



**Plasma-Assisted Waste-to-Energy Processes. Figure 3**  
Effect of constituent materials and moisture on heating value of MSW [5]

the syngas, and lower emissions, as we will see later on. Figure 5 below shows that subjecting the products of gasification to plasma treatment creates a higher quality syngas by increasing the  $H_2$  to  $CO$  ratio.

### Non-transferred Arc Plasma Torch

Non-transferred arc plasma torches are commonly used in the treatment of wastes. Electricity is transformed into thermal energy by means of an electric discharge between the cathode and the anode contained in a water-cooled torch. This device can be used either with hot (thermo-ionic) rode-type cathode (Fig. 6) and cold tubular water-cooled cathode.

Figure 7 is a schematic of the Europlasma non-transferred plasma torch that utilizes a cold cathode:

### Transferred Plasma Arc

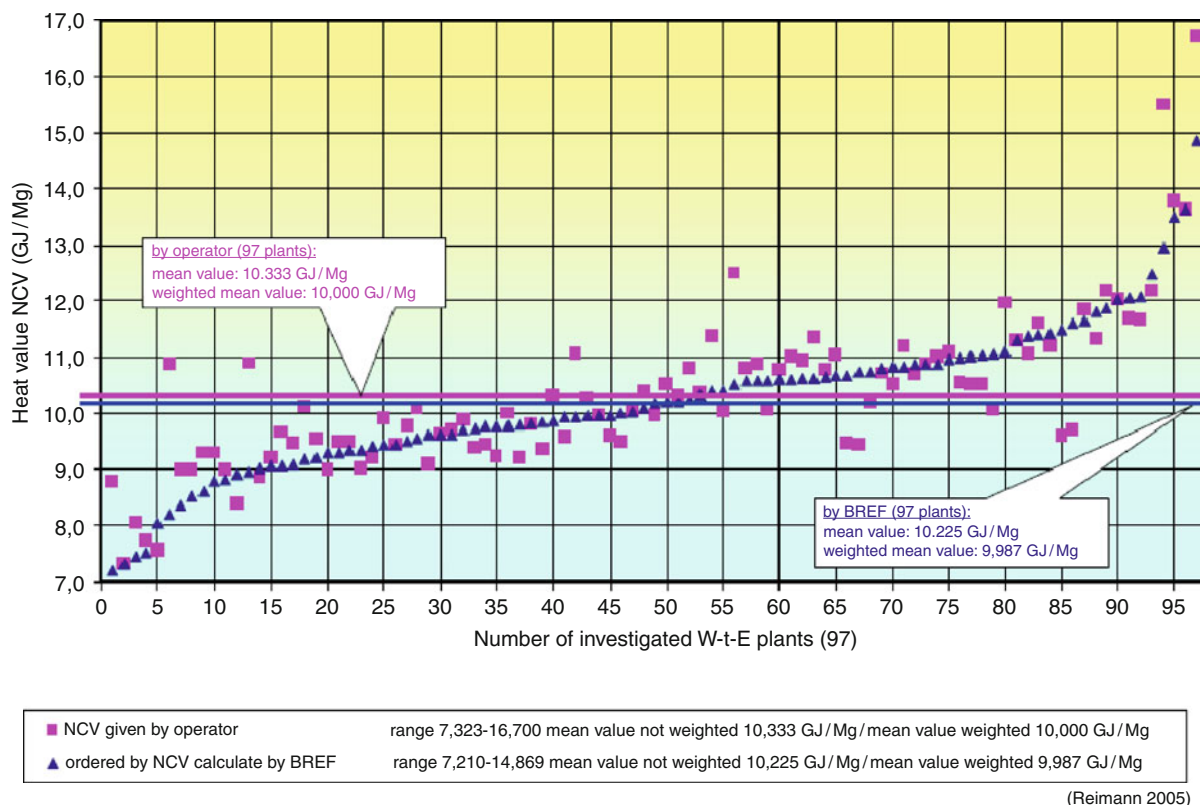
In the case of transferred arc, one of the electrodes is external to the torch. The electricity flows through the gas column issuing from the torch into the molten metal or slag below the torch, which is connected to

the external electrode (Fig. 8). The peak temperature of the arc plasma can range from 12,000 to over 20,000 K.

Since the plasma is produced outside of the water-cooled body of the torch, this device is thermally more efficient than the non-transferred arc torch. The cathode can be either a water-cooled metal tube or non-cooled graphite tube that is consumed slowly by sublimation. In this case, the thermal loss is reduced, but the cathode needs to be replenished. The anode is generally made from a high thermal conductivity metal that is water cooled at the outer end so as to avoid melting it.

### Energy and Material Balances in Plasma-Assisted Gasification of MSW

As described earlier, plasma-assisted gasification volatilizes MSW in an oxygen-deficient environment where the waste materials are decomposed and partially oxidized to the basic molecules of  $CO$ ,  $H_2$ ,  $CO_2$  and  $H_2O$ . Thus, the organic fraction of the waste is converted into a synthesis gas ("syngas") that contains most of the chemical energy of the waste. Also, the inorganic fraction of the MSW can be converted into an inert vitrified

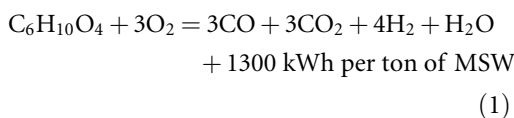


**Plasma-Assisted Waste-to-Energy Processes. Figure 4**  
Calorific values of MSW of 97 E.U. WTE facilities [7]

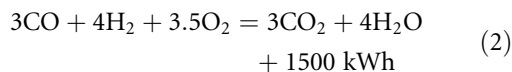
glass so that there is no ash remaining to be landfilled. Furthermore, the plasma reactor can treat all waste materials, as the only variable is the amount of energy needed to melt the waste. Any kind of feedstock, other than nuclear waste, can be directly processed. Controlling the temperature of the output gasses by modifying the temperature allows for better control of the syngas composition.

For a typical MSW of total calorific value of 2,800 kWh/t, the two steps of the overall process can be represented by the following chemical equations:

- Gasification by means of partial combustion with oxygen (assuming zero reactor heat loss):

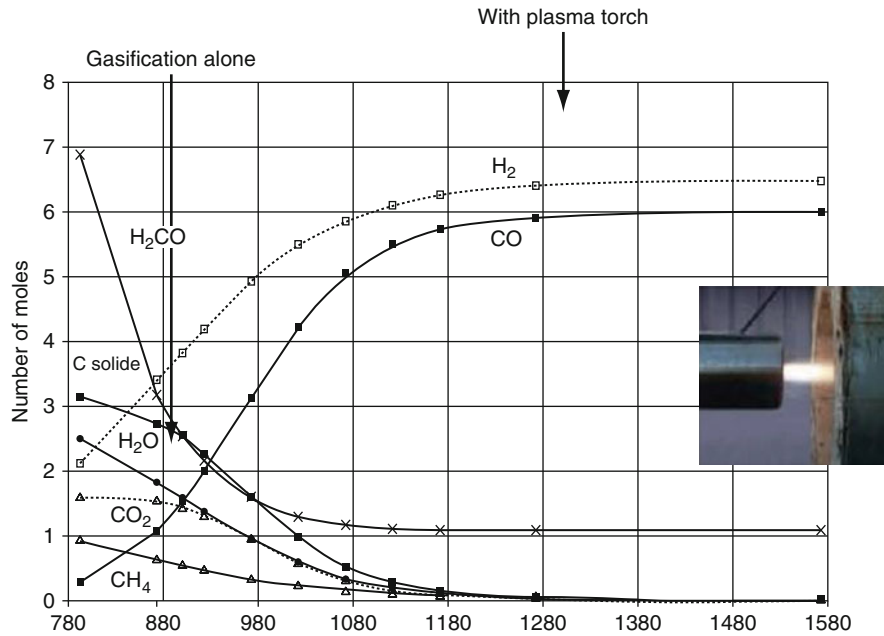


- Gas turbine combustion (assuming zero turbine heat loss):



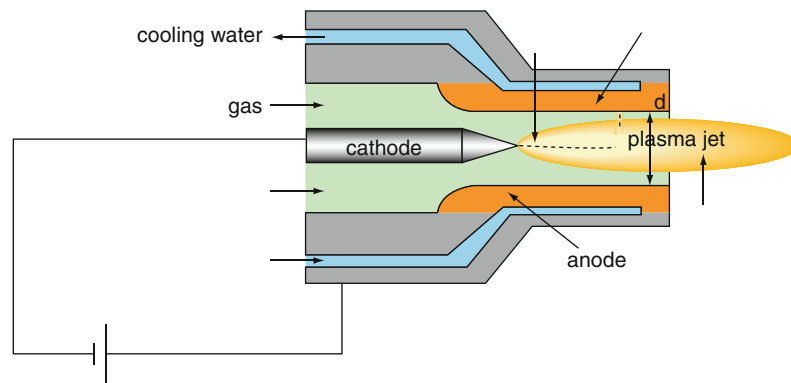
Typically, the syngas produced in plasma-assisted gasification has about 30% of the heating value of natural gas.

At an assumed thermal efficiency of 50% of the gas turbine, the electricity generated will be equal to  $1,500 \text{ kWh} \times 50\% = 750 \text{ kWh/t}$  of solid wastes. The oxygen required for partial combustion should be provided in the form of industrial oxygen, to avoid the introduction of about four parts of nitrogen per part of oxygen. The production of 1 t of industrial oxygen (95%  $O_2$ ) requires about 250 kWh of electricity. Since the gasification reaction (1) requires three moles of oxygen per mole of combustibles, the gasification of 1 t of MSW containing 20% of moisture and 20% of inorganic materials will require



Plasma-Assisted Waste-to-Energy Processes. Figure 5

Effect of plasma on products of gasification [1]



Plasma-Assisted Waste-to-Energy Processes. Figure 6

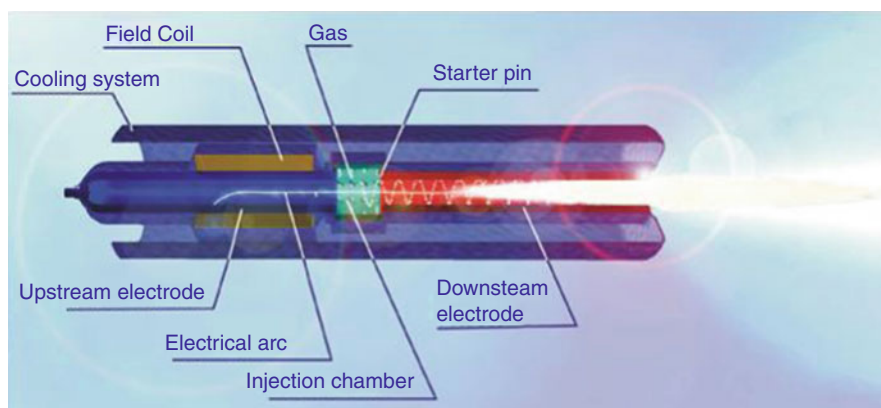
Non-transferred arc plasma torch [8]

$$1000 \times 60\% / 146 \times 96 / 1000 \times 250 \text{ kWh} \\ = 75 \text{ kWh of electricity per ton of MSW processed.}$$

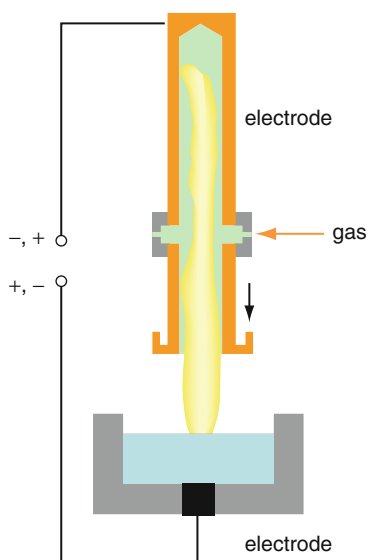
Thermodynamic considerations show that gasification of various types of waste will yield the syngas compositions shown in Table 1. Figure 9 shows the composition of syngas produced by the gasification of a typical MSW.

### Plasma-Assisted Processes for Treating MSW

Plasma processes have been used widely for the destruction of asbestos and other hazardous wastes. However, due to their high consumption of electricity, sole use of plasma energy is not viable economically for the treatment of low-value materials, such as MSW. However, *plasma-assisted* gasification processes are being developed and may offer environmental and economic benefits.



**Plasma-Assisted Waste-to-Energy Processes. Figure 7**  
Europlasma non-transferred arc DC plasma torch [9]



**Plasma-Assisted Waste-to-Energy Processes. Figure 8**  
Transferred arc plasma torch [8]

There are several plasma-assisted gasification technologies where plasma torches are used to accelerate the gasification process, to crack the product of volatilization to CO and H<sub>2</sub>, and to vitrify the inorganic component of MSW. The processes to be described in this essay are the gasification process of Alter NRG that is based on the Westinghouse Plasma Technology, and the plasma-assisted process developed by Europlasma. The potential main advantages of plasma-assisted processes, as compared to conventional WTE plants, are

the reduction of exhaust gas flow rate, an overall installation with smaller footprint because of more compact equipment, lower capital investment for a given throughput, and faster start-up and shutdown times.

#### The Alter NRG Westinghouse Plasma Corporation Process

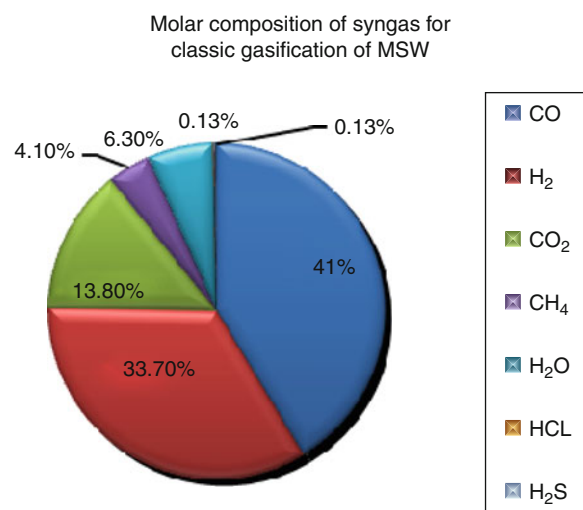
In 2006, Alter NRG acquired the Westinghouse Plasma Corporation (WPC), a leading plasma gasification technology. The non-transferred plasma torch consists of a pair of tubular water-cooled copper electrodes, the operating gas being introduced through an annular space between the electrodes. A schematic diagram of the operation of the torch is shown in Fig. 10. Figure 11 shows two views of the largest plasma torch of Alter NRG that has an operating range of 80–500 kW (Marc 11 plasma torch).

The WPC torches have been used extensively in metal melting cupolas, but their most important applications have been in the destruction of hazardous waste and the vitrification of WTE ash, mostly in Japan. Since the WPC torch is water-cooled, the efficiency of converting electricity into heat ranges from 60% to 75%.

The MSW gasification process developed by Alter NRG is based on a cupola furnace fired by the WPC plasma torches (Fig. 12). This technology is well proven and currently used in several processing plants in Japan. Alter NRG has tested and offers this process for the gasification of MSW, biomass, petroleum coke, and hazardous wastes to produce syngas.

**Plasma-Assisted Waste-to-Energy Processes. Table 1** % Molar (volume) composition of the syngas from different feedstocks

	CO	H <sub>2</sub>	CO <sub>2</sub>	CH <sub>4</sub>	H <sub>2</sub> O	HCL	H <sub>2</sub> S
MSW (typical)	41.0	33.7	13.8	4.1	6.3	0.13	0.13
Carpet	33.2	43.1	6.8	8.8	4.9	0.02	0.03
Tire	56.9	18.9	1.5	22.2	0.3	0.04	0.00
Biomass	27.5	36.1	20.1	1.4	14.7	0.03	0.00
Med waste	27.9	37.8	18.2	1.8	13.7	0.03	0.65
ASR	29.8	37.4	17.3	2.1	12.0	0.00	0.64
Oil	48.8	25.6	2.2	21.1	0.6	1.61	0.00
Bituminous	55.9	23.9	4.1	12.8	1.0	1.71	0.00



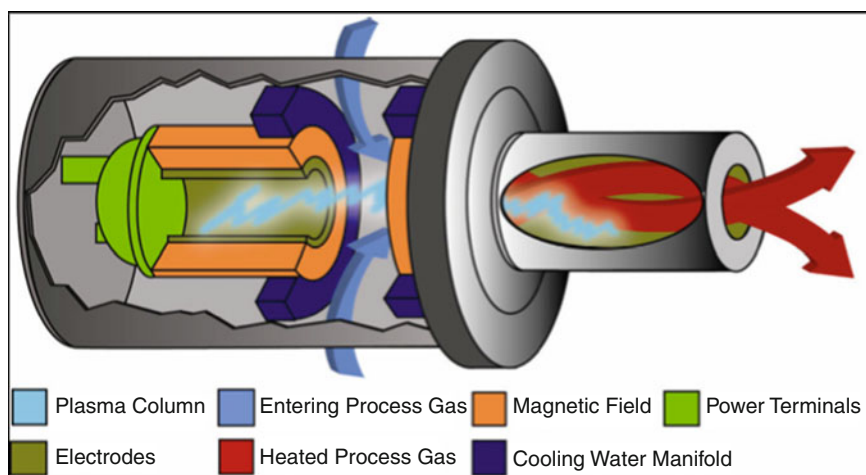
**Plasma-Assisted Waste-to-Energy Processes. Figure 9** Composition of syngas from gasification of typical MSW

The Alter NRG-WPC process uses up to six plasma torches at the bottom part of the gasifier. A bed of coke is formed within the cupola using metallurgical coke (“met coke”) to absorb and retain the heat energy from the plasma torches and provide a “skeleton” that supports the MSW feed as it descends through the gasification reactor and is converted to gas and liquid slag; this action is similar to the phenomena occurring in an iron blast furnace. The met coke and the MSW are fed from the top of the refractory-lined gasification vessel. Figure 13 is a photograph of the Alter NRG pilot reactor at Madison, Pennsylvania, USA.

In 2010, Alter NRG was using its industrial-size pilot plant at Madison, PA, to gasify wood chips to syngas that is stored in large gas tanks and is then used by another company, located next to the Alter NRG plant, to produce ethanol. There were four industrial plants using the Alter NRG gasifier: two in Japan (one on MSW plus automobile shredder residue; the other on MSW and wastewater sludge) and two in India (Pune and Nagpur) processing hazardous waste. All these plants use the smaller Marc 3 torches (300 kW capacity), quench and clean the syngas, and then combust it with air to generate steam. The largest plant, in Japan, has a nominal capacity of 300 t MSW per day, while the Indian plants are of 72 t/day capacity.

Operating experience has shown that the electrodes of the Marc 3 torch have a lifetime of up to 500 h. Used electrodes are repaired and reused. The largest WPC torch, Marc 11 is currently used in Quebec for metal smelting. The lifetime of this torch is up to 1,200 h. Six Marc 11 torches will be required for a 750 t/day plant processing MSW.

This process can handle any moisture content in the MSW since water is vaporized along with the syngas. However, the feedstock must be less than 25 cm in size to facilitate feeding into the furnace. The process is controlled by maintaining the temperature of the gas exiting the gasifier between 1,000°C to 1,100°C. At the bottom of the cupola, the inorganic components in the MSW are melted into a slag layer and a metal layer underneath the slag. These liquids are tapped intermittently from the furnace.



**Plasma-Assisted Waste-to-Energy Processes. Figure 10**

The WPC non-transferred arc plasma torch [10]



**Plasma-Assisted Waste-to-Energy Processes. Figure 11**

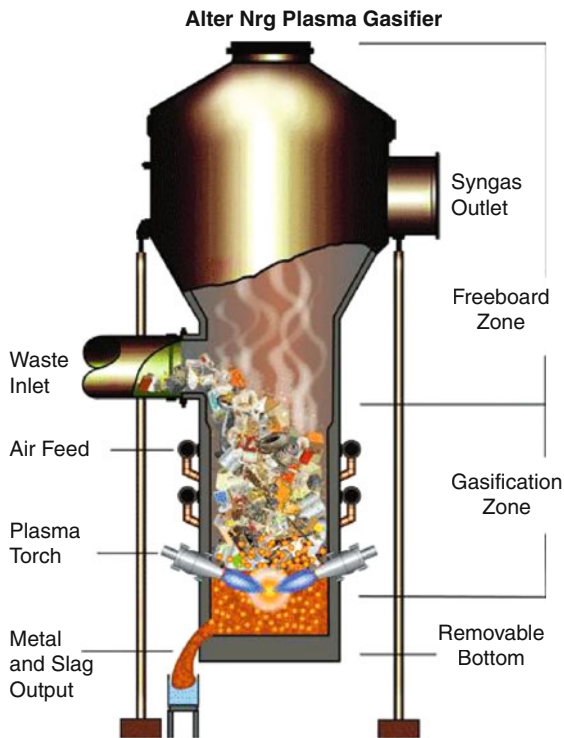
Side and front view of the Alter NRG Marc 11 plasma torch

The plasma torches are controlled independently of each other, and a torch can be removed for maintenance while the furnace is operating. The gasifier is working at a slightly negative pressure to avoid gaseous leaks. There is a small gap between the torches and the furnace wall so that a small amount of air infiltrates into the furnace (Fig. 14).

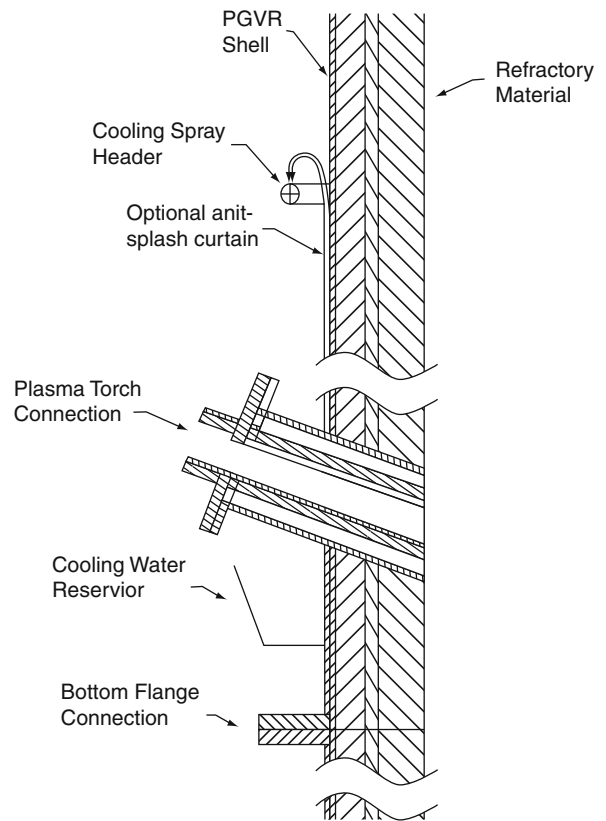
As in the case of other gasification processes, the syngas produced contains about one third of the energy content of natural gas. Therefore, the gas turbine used to generate electricity has to be compatible with a lower energy gas. The ultimate goal of Alter NRG is to operate with MSW feed plus 4% of met coke and generate power by means of the Integrated Gasification Combined Cycle (IGCC, Fig. 15).

The largest project for plasma gasification of MSW was announced in 2006 as a partnership between Alter NRG and Geoplasma, at St Lucie, Florida. The initial plan was to construct a plant processing 1 million tons of waste per year. However, due to the lack of investors and public opposition, the project was scaled down to a 500-t/day plant (about 150,000 t/year). This plant will consist of two lines of total nominal capacity of 500 t/day, but maybe increased to 750 t/day. The projected met coke and limestone use will be 4% and 7.9%, respectively. The input materials to the gasification reactor are shown in Fig. 16, and Fig. 17 is a schematic flowsheet of this potential application.

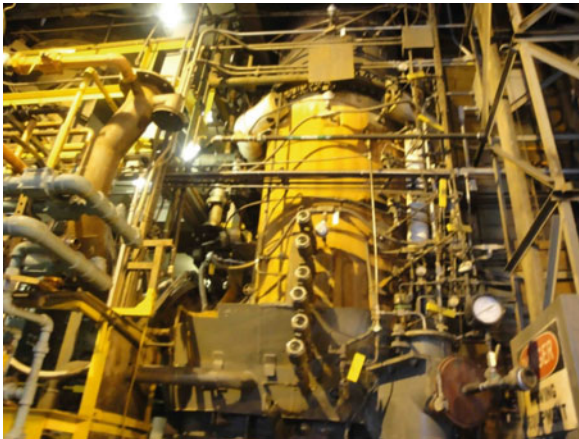
An overall energy balance for such a system was calculated by Caroline Ducharme (Ducharme 2010).



Plasma-Assisted Waste-to-Energy Processes. Figure 12  
The Alter NRG plasma gasifier [10]



Plasma-Assisted Waste-to-Energy Processes. Figure 14  
Schematic diagram showing how torch is introduced  
through furnace wall



Plasma-Assisted Waste-to-Energy Processes. Figure 13  
The Alter NRG gasification reactor viewed from the bottom [4]

The electricity needed to shred the MSW was estimated at less than 10 kWh/t. As noted earlier, a typical MSW has a calorific value of about 10 MJ/kg which corresponds to 2,800 kWh/t. The energy contained in the

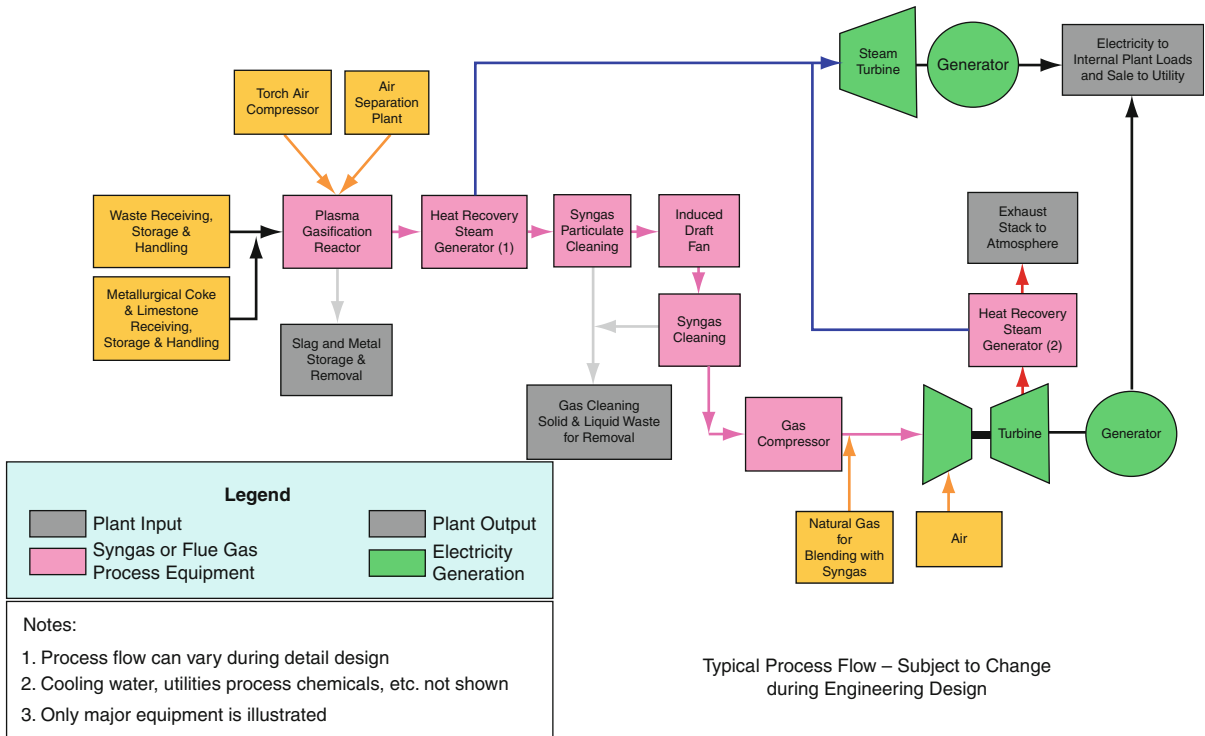
metallurgical coke (LHV: 32.8 MJ/kg) was calculated from

$$\begin{aligned} \text{One ton of MSW} &\times 4\% \times 32.8 \text{ MJ/kg} \times 1000 \text{ kg/ton} \\ &= 1312 \text{ MJ or } 335 \text{ kWh/ton MSW} \end{aligned}$$

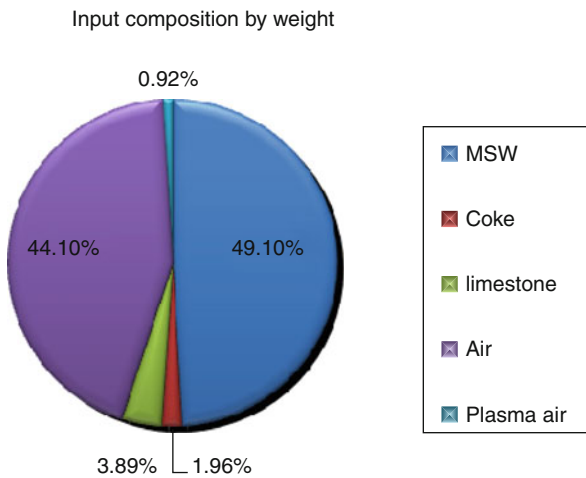
The thermal energy provided by six 600-kW torches used in the Alter NRG reactor is  $6 \times 600 \times 75\% = 2,700$  kWh (the plasma torches need to be water-cooled so that their average efficiency of converting electricity to heat is assumed to be 75%). The electricity consumed by the six plasma torches in a plant of 750 t/day (31.25 t/h) would be 3,600 kWh/h, corresponding to about 115 kWh of electricity per ton MSW processed. Table 2 shows the distribution of energy inputs to the gasification plant.

The energy outputs are the heat loss from the reactors, the heat loss in the cooling water of the torches, the heat carryover in the vitrified ash, and the chemical





Plasma-Assisted Waste-to-Energy Processes. Figure 15 Integrated Gasification Combined cycle [10]

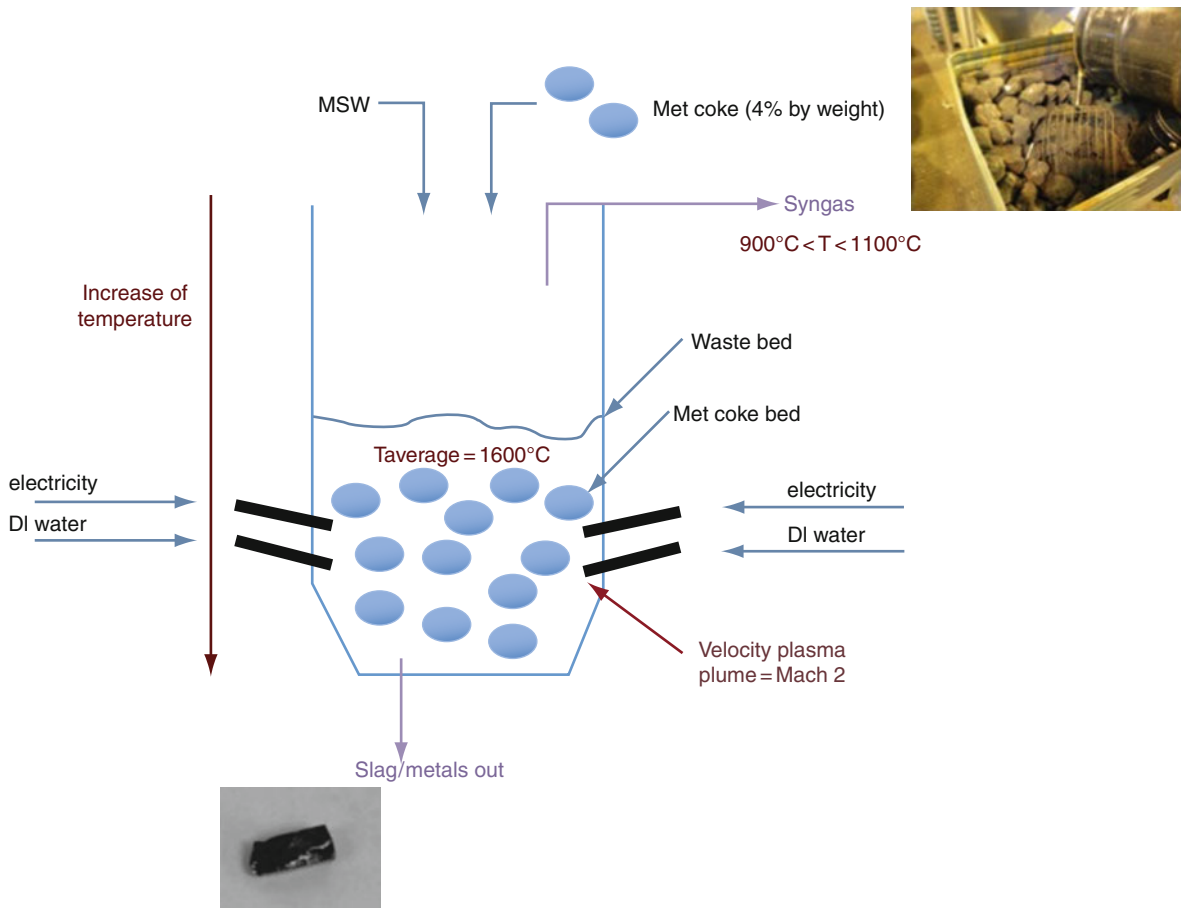


Plasma-Assisted Waste-to-Energy Processes. Figure 16 Input composition to WPC reactor by weight [10]

plus sensible heat in the syngas product. Assuming a 10% heat loss from the gasification reactor plus the water-cooling system of the torches and the vitrified ash, the syngas should carry 90% of the energy input by

the MSW, coke, and plasma torches. An estimated 80% of this energy is in the form of chemical energy in the syngas and 20% is thermal energy, in the form of sensible heat. When the syngas is quenched, as in the present Alter NRG process, the sensible heat is not recovered. Therefore, the chemical energy content in the syngas will be:  $0.90 \times 0.80 \times 3,136 = 2,258$  kWh. If the syngas is used to power a gas turbine at 45% efficiency, the gross electrical energy generated will be 1,015 kWh of electricity per ton of MSW. However, some of this energy must be used in the operation of the plant, i.e., shredding of MSW, production of industrial oxygen for combustion, operation of plasma torches, and all other uses of electricity within the plant (Table 2). The consumption of electricity for oxygen production was estimated earlier at 75 kWh/t of MSW.

- Shredding of MSW: 10 kWh/t MSW
- Operation of the Air Separation Unit: as per earlier discussion, an estimated 75 kWh of electricity would be used per ton of MSW processed.



**Plasma-Assisted Waste-to-Energy Processes. Figure 17**  
Flows of materials and energy into the WPC reactor [4]

- Operation of the plasma torches: 115 kWh/t of MSW processed
- All other needs of plant (assumed to be 75% of those of a conventional WTE plant that has to clean a much larger volume of gas): 75 kWh

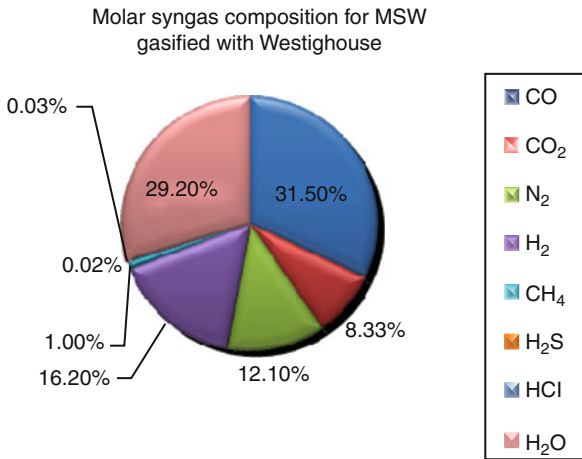
By subtracting the above internal uses of electricity from the 1,015 kWh generated by the syngas turbine yields the net electricity generated by this plant per ton of MSW processed: 740 kWh. This calculated number is somewhat higher than a conventional grate combustion WTE of the same size that generates 650 kWh per metric ton of MSW containing 2,800 kWh of chemical heat. [Figure 18](#) shows the projected composition of syngas produced by the WPC gasification of MSW.

**Plasma-Assisted Waste-to-Energy Processes. Table 2**  
Energy inputs to Alter NRG plant, per ton of MSW processed

Inputs	In kWh	In % of total inputs
MSW	2,800	85.8
Met coke	335.8	10.3
Energy from torches	115.2	3.5
Total	3,251	100

### The Europlasma WTE Process

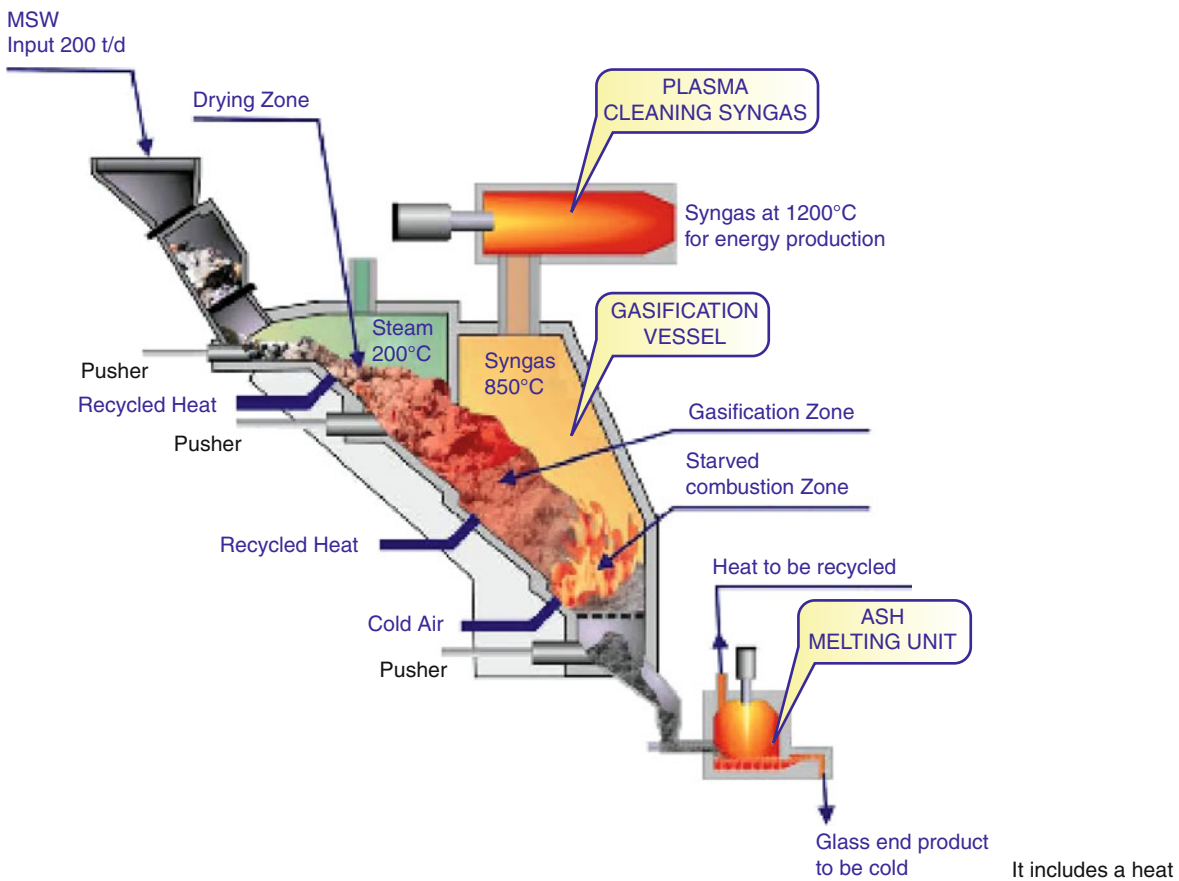
Europlasma is a French company and one of the world leaders in plasma technology as applied to the thermal treatment of wastes. They have been very successful in



Plasma-Assisted Waste-to-Energy Processes. Figure 18 Syngas composition for MSW gasified by the WPC process

using non-transferred arc torches for vitrifying incinerator residues and have also developed a process for treating asbestos contaminated wastes. The plasma torch consists of two tubular, coaxial, water-cooled, copper electrodes separated by a tubular gap through which flows the plasma forming gas. The thermal efficiency of the Europlasma torches is in the order of 75–80%. Europlasma has also developed a special plasma torch for cracking the gasification syngas, called “TurboPlasma.”

Figure 19 shows the Europlasma gasification process. It includes a stoker grate auto-thermal gasifier, a plasma-fired chamber for cracking the gasification gas to hydrogen and carbon monoxide, and a second plasma torch for vitrifying the solid product of gasification; the gas



Plasma-Assisted Waste-to-Energy Processes. Figure 19 The Europlasma process [9]

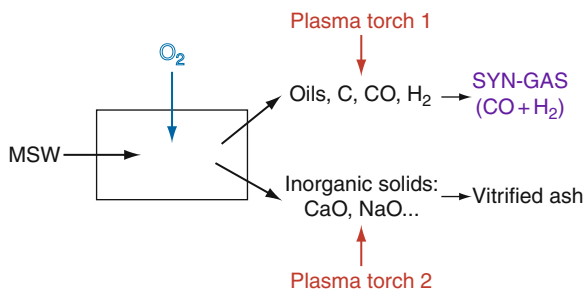
flow from the vitrification unit flows into the gasification unit and provides the required heat for gasification. The clean syngas may pass through a heat exchanger to recover its sensible heat and then through a scrubber to remove acid gasses and particulate matter; or quenched directly by scrubbing. Figure 20 is a schematic diagram of the processes, such as Europlasma, that use two torches for syngas cracking and ash vitrification.

Europlasma is currently constructing their first MSW gasification plant at Morcenx, France; start up is planned for end of 2011. The plant capacity will be 50,000 t of waste per year.

Since there are no data as yet from the Morcenx plant of Europlasma in France, Ducharme [4] relied on an energy analysis conducted by Sunbeam for Credit Suisse of a proposal to build a Europlasma plant in New Jersey. This plant was to process 400 t/day of MSW mixed with 3% shredded tires. The assumed plant availability was 90% corresponding to a nominal capacity of 120,000 t per year. The LHV of the MSW was 2,800 kWh/t and of the shredded tires 9,690 kWh/t. Therefore, the MSW-tire mix had an average LHV of 3,090 kWh/t.

According to Sunbeam, the proposed 400 t/day plant would use 4,800 kW for the plasma torches distributed as follows: 4,000 kW (83%) for the syngas polishing torch and 800 kWh (17%) for the ash vitrification torch. Thus, according to the Sunbeam data, the electricity consumption per ton processed would be:

$$4800 \text{ kW} \times 24/400 = 288 \text{ kWh per ton of MSW}$$



**Plasma-Assisted Waste-to-Energy Processes. Figure 20**  
Flowsheet of plasma assisted-gasification process

On a per ton MSW basis, this projected value is more than double the electricity consumption of the Alter NRG plant and, therefore, questionable. Therefore, it is necessary to wait for operating data after the start-up of the Europlasma plant at Morcenx.

## Environmental Impacts

The first noticeable difference of plasma-assisted processes from classic grate combustion is that the syngas is cleaned before combustion, which should be less costly than post-combustion cleaning of WTE flue gas. The final emissions of a plasma-assisted process will depend on the level of cleaning of the syngas, with the exception of NO<sub>x</sub> that will not be formed during the gasification process. However, some NO<sub>x</sub> will be formed during combustion in the power generation equipment. Dioxins and furans can be avoided due the high heat of the plasma treatment, but they can form “de-novo” during the cooling of the syngas. However, the dioxin emissions of modern grate combustion WTE plants are so low (less than 0.5 g TEQ per million tons of MSW) that they are insignificant.

In contrast to conventional grate combustion that has no liquid effluents, quenching of the syngas results in an aqueous stream that must be cleaned before discharging.

A definite advantage of plasma-assisted processes is that the vitrified slag is impervious to leaching and can definitely be used for construction.

## Future Directions

Plasma torches have been highly developed and are an excellent tool for converting electricity to an extremely high temperature gas. As described in this essay, plasma torches are used for thermally treating hazardous materials such as asbestos and can be used, in combination with partial combustion, for treating any type of solid wastes, including MSW. Such plasma-assisted WTE processes are in operation in Japan and India and an industrial plant is under construction in France. The advantages they offer over conventional grate combustion are a much reduced volume of process gas to clean and the potential of higher thermal efficiency in using

the syngas in a gas turbine, rather than generating steam for a steam turbine, as in the case of conventional grate combustion.

As mentioned in other sections of this Encyclopedia, the major cost factor of thermally treating 1 t of MSW is the repayment of the capital cost. This is where plasma-assisted WTE and other gasification processes can compete with grate combustion and also widen the application of WTE over landfilling: by offering gasification plants that, because of their compactness and higher rates of reaction, will be less costly to build than the giant WTE plants that are based on grate combustion.

With regard to higher energy production, plasma-assisted WTE processes, must “invest” electricity in the operation of the plasma torches. Therefore, it is preferable to use the syngas in a gas turbine that offers higher thermal efficiency than in steam turbines such as those are used by grate combustion processes. An analysis of several plasma-assisted WTE processes at different stages of commercialization by Ducharme [4] showed that most are quoting numbers of electricity generation (e.g., 1,000 kWh/t MSW) that are much higher than the numbers calculated from material and energy balances. Also, the produced syngas has a calorific value equal to one third of natural gas. To overcome this problem, the developing companies have two options, either to blend syngas with natural gas or to use specially adapted turbines.

In conclusion, plasma-assisted gasification of solid wastes is a very interesting process with potential for future application. First, using a reducing atmosphere and producing a relatively smaller amount of process gas facilitates the gas cleaning system. Second, controlling the amount of heat input to the process by means of the plasma torches allows controlling the composition of the syngas. The hydrogen to carbon monoxide ratio can be modified easily, according to the needs of the user. The next decade will show how plasma-assisted gasification of MSW evolves. The plants under planning or construction should provide reliable information on capital and operating costs per ton of solids treated and this technology may provide an alternate route for the thermal treatment of MSW.

## Bibliography

### Primary Literature

1. Fauchais P, Vardelle A (1997) Thermal plasmas. *IEEE T Plasma Sci* 5(6):1258–1280
2. Kogelschatz U (2004) Atmospheric-pressure plasma technology. *Plasma Phys Control Fusion* 46:B63–B75, s.I
3. Heberlein J, Murphy AB (2008) Thermal plasma waste treatment. *J Phys D Appl Phys* 41:s.I., 053001
4. Caroline D (2010). Technical and economic analysis of plasma-assisted waste-to-energy processes. M.S. Thesis, Earth and Environmental Engineering, Columbia University. [www.seas.columbia.edu/earth/wtert/sofos/ducharme\\_thesis.pdf](http://www.seas.columbia.edu/earth/wtert/sofos/ducharme_thesis.pdf)
5. Themelis NJ, Kim YH (2002) Energy recovery from New York City waste. *Waste Manag Res* 20:223–233, s.I
6. HSC. Chemistry, 2010. <http://www.hsc-chemistry.com>
7. Reimann DO (2008). CEWEP energy report (Status 2001–2004). [www.cewep.org](http://www.cewep.org)
8. NOVELECT (2003). Innovative applications of thermal plasmas (in French). s.l.: EDF publication
9. Europlasma. <http://www.europlasma.com/>
10. Corp, ALter NRG/ Westinghouse Plasma (2010). [www.alternrg.ca](http://www.alternrg.ca)
11. Igiorno V, De Feo G, Della Rocca C, Napoli R (2003) Energy from gasification of solid wastes. *Waste Manag* 23:1–15

### Books and Reviews

- Alexander F (2008) Plasma chemistry. Cambridge University Press, New York
- Boulos MI, Fauchais P (1994) Thermal plasmas, fundamentals and applications, 1st edn. Springer, Dordrecht, p 468
- Bridgwater AV (1995) The technical and economic feasibility of biomass gasification for power generation. *Fuel* 74(5): 631–653
- Bridgwater AV, Toft AJ, Brammer JG (2002) A techno-economic comparison of power production by biomass fast pyrolysis with gasification and combustion. *Renew Sust Energy Rev* 6:181–248
- Clark BJ, Rogoff MJ (2010). Economic feasibility of a plasma gasification plant. Proceedings of the 18th annual North American waste to energy conference (NAWTEC 18–35), City of Marion, Iowa, 11–13 May 2010
- Gomez E, Amutha Rani D, Cheeseman CR, Deegan D, Wisc M, Boccaccini AR (2009) Thermal plasma technology for the treatment of wastes: a critical review. *J Hazard Mater* 161:614–626
- Hackett C, Williams RB, Durbin TD, Welch W, Pence J, Jenkins BM, Aldas R, Salour D (2004) Evaluation of conversion technology processes and products, University of California
- Heberlein J (2002) New approaches in thermal plasma technology. *Pure Appl Chem* 74(3):327–335

- Juniper Consulting (2008) Independent waste technology report, the Alter NRG/Westinghouse plasma gasification process
- Klein A, Themelis NJ (2003) Energy recovery from municipal solid wastes by gasification. North American waste to energy conference (NAWTEC 11) 11 proceedings, ASME International, Tampa, FL
- Kogelschatz U (2004) Atmospheric-pressure plasma technology. *Plasma Phys Control Fusion* 46:B63–B75
- Murphy AB, McAllister T (2001) Modeling of the physics and chemistry of thermal plasma waste destruction. *Phys Plasmas* 8:2565–2572
- Niessen WR, Markes CH, Sommerlad RE (1996) Evaluation of gasification and novel thermal processes for the treatment of municipal solid waste. NREL/TP-430-21612, Aug 1996
- Plasco Energy Group. <http://www.plascoenergygroup.com/>
- Solonenko OP. Thermal plasma torches and technologies. Cambridge Science International Publishing, Cambridge
- Titus CH, Surma JE (1998). Integrated environmental technologies, LLC, enhanced tunable plasma-melter vitrification systems. Patent number 5,811,752, 22 Sep 1998
- Willis KP, Osada S, Willerton KL (2010) Plasma gasification: lessons from ecovalley WTE facility. Proceedings of the 18th annual North American waste to energy conference (NAWTEC 18–3515), Orlando, FL, 11–13 May 2010

- CS** Charge sustaining
- DoD** Depth of discharge
- EV** Electric vehicle
- HEV** Hybrid electric vehicle
- PHEV** Plug-in hybrid electric vehicle
- SOC** State-of-charge

### Definition of the Subject

The plug-in hybrid electric vehicle (PHEV) is vehicle that has the capability of accepting part of its propulsion energy from the electric utility grid. However, like the conventional hybrid electric vehicles, it also draws part of its traction energy from its fuel tank. This will afford the plug-in hybrids dual fuel flexibility, electric charge, and fossil fuel. Depending on the design of the vehicle drive train and its operation, the plug-in hybrid vehicle can behave as pure electric, pure engine, or hybrid vehicle.

### Introduction

It is now well recognized that the hybrid electric vehicle (HEV) is much more efficient and cleaner than the vehicle powered by gasoline and diesel engine alone [1]. The HEV also has high vehicle performance and more user acceptability than pure battery powered electric vehicle (EV). However, all of the HEV still comes from burning fossil fuel, gasoline, or diesel. On the other hand, the EV has certain advantages over HEV, mostly zero emission, independence from petroleum, and perhaps low operating cost. However, the major disadvantage of EV is the range limitation and long battery charging time. It should be noted that only fraction of battery energy is used in a conventional HEV, in which the variation of the battery state of charge (SOC) is limited to a narrow band [1]. This fact implies that most of the battery energy just stays there unused.

Using most of the battery energy, drawn from the utility grid, to displace part of petroleum fuel is the major advantage of plug-in hybrid electric vehicles. The plug-in hybrid electric vehicle (PHEV) is a form of hybrid electric vehicle, which is specifically designed and operated to fully use the energy in the battery for period of distance of pure EV operation. It has the advantage of being both pure EV and an HEV. When the battery is in high charge state, after overnight

## Plug-in Hybrid Electric Vehicles

MEHRDAD EHSANI

Department of Electrical & Computer Engineering,  
Texas A&M University, College Station, TX, USA

### Article Outline

- Glossary
- Definition of the Subject
- Introduction
- Statistics of Daily Driving Distance
- Motor and Battery Power
- Energy Consumption in Typical Driving Cycles
- Operation Strategies
- Battery Design
- Summary and Future Directions
- Bibliography

### Glossary

- AER** All-electric range
- CD** Charge depletion
- CDR** Charge-depletion range

charging, the vehicle may be operated in pure EV mode: charge-depletion (CD) mode. In this mode, all the traction energy is supplied by the battery. When the state of charge of the battery reaches a certain low level, such as 30%, depending on the characteristics of the battery, the engine is started and the vehicle goes to battery charge-sustaining (CS) mode. In the CS mode, the energy in the battery is maintained around this level till the end of the trip, at which time the battery is charged to its full state through utility grid [2–6].

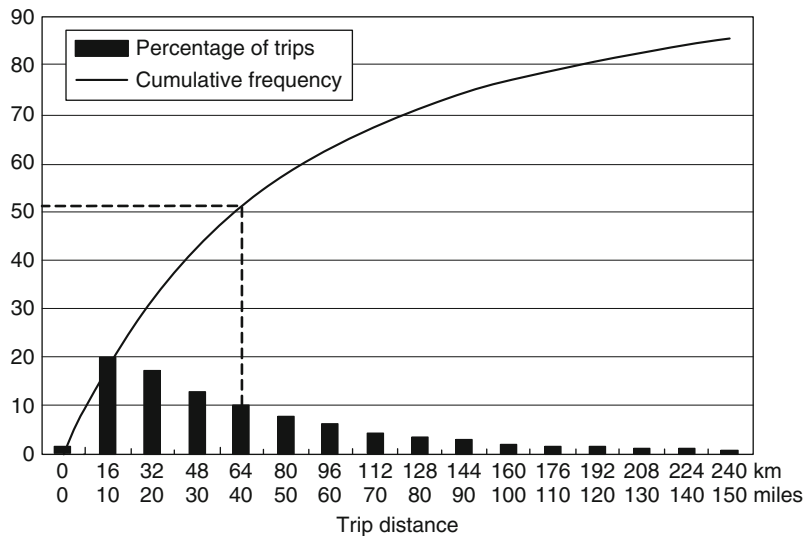
Thus, the PHEV shares the combined operation characteristics of pure EV and CS HEV. However, the advantages of PHEV over pure EV and HEV are (1) displacement of significant amount petroleum fuel by electric energy that comes from utility grid, (2) longer range than a pure EV, (3) no fuel consumption and emission during pure EV mode, (4) and lower overall fuel consumption and emission.

**Statistics of Daily Driving Distance**

As mentioned above, using the energy stored in the battery from the utility grid to displace part of petroleum fuel is the major feature of the plug-in hybrid electric vehicles. The fraction of the petroleum fuel displaced by electricity depends mostly upon the amount of electrical energy drawn from the utility grid. That is, the energy capacity of the battery and

the total driving distance, which is usually the diurnal driving distance, and electrical traction power usage profiles which is related to the drive cycle features and vehicle control strategies. For optimal PHEV design, especially the battery size, understanding of diurnal driving distance in some detail is very helpful.

Figure 1 shows a histogram of diurnal driving distance distribution and their cumulative frequency in 1995, from National Personal Transportation Survey (NPTS) data [2]. The cumulative frequency or utility factor in Fig.1 represents the percentages of the total driving time (days) during which the daily driving distances are less than or equal to the said distance on the horizontal axis. Figure 1 reveals the fact that about half of daily driving distance is less than 40 miles (64 km). If a vehicle is designed to have 40 miles (64 km) of pure EV range, that vehicle will have half of its total driving distance in pure EV mode. Even if the daily traveling distance is beyond 40 miles, there is still a significant amount of petroleum fuel that can be displaced by electricity, due to the pure EV driving taking a large portion of the daily traveling. Research also shows that even if the pure EV range is less than 40 miles, such as 20 miles (32 km), the portion of petroleum fuel that can be displaced in normal daily driving is still very significant [2].



**Plug-in Hybrid Electric Vehicles. Figure 1**  
Diurnal driving distance distribution and cumulative factor

### Motor and Battery Power

A PHEV consists of a considerable amount of pure EV operation in which the electric motor and the battery are the sole power plant and the energy source for the vehicle. The motor and battery power must be large enough to stratify the peak power requirement of vehicle. Otherwise, the vehicle cannot accomplish the demanded driving cycle.

The traction power of a vehicle, measured on the drive wheels, can be expressed as:

$$P_t = \frac{V}{1000} \left( Mg f_r + \frac{1}{2} \rho_a C_D A_f V^2 + M \delta \frac{dV}{dt} + M g i \right) (kW) \quad (1)$$

where  $M$  is the vehicle mass in kg,  $V$  is vehicle speed in m/s,  $g$  is gravity acceleration,  $9.81 \text{ m/s}^2$ ,  $\rho_a$  is the air mass density,  $1.205 \text{ kg/m}^3$ ,  $C_D$  is the aerodynamic drag coefficient of the vehicle,  $A_f$  is the front area of the vehicle in  $\text{m}^2$ ,  $\delta$  is the rotational inertia factor,  $dV/dt$  is the acceleration in  $\text{m/s}^2$ , and  $i$  is the grade of road. In standard driving cycles, road is flat with  $i = 0$ .

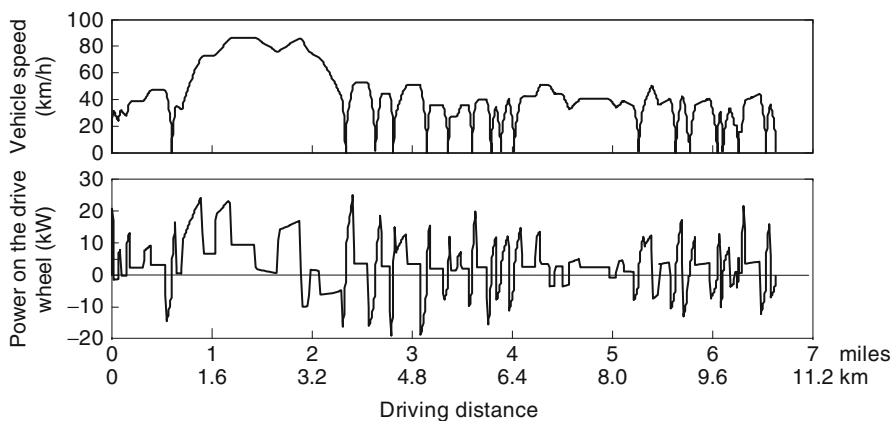
Figure 2 is a diagram showing vehicle speed and traction power measured on the drive wheels versus the travel distance in the FTP 75 urban driving cycle. The vehicle design parameters used in the computation are listed in Table 1.

Figure 2 indicates that the peaking traction power on the drive wheels is around 25 kW. However, there are power losses from the battery to the drive wheels. In order to meet the drive cycle power requirement, the

motor output power should be designed to include the power losses from the motor shaft to the drive wheels. Suppose the efficiency from motor shaft to the drive wheels is 90%, then the motor shaft power rating is round 28 kW. It should be noted that this motor power required is related to the vehicle speed at which this peak power occurs. For example, the peaking power in Fig. 2 occurs at the vehicle speed of 50 km/h (31.25 mph). In the motor power design, one must be sure that the motor can produce this peak power at this vehicle speed. Similarly, the peaking power of the battery should include the losses in the electric motor, power electronics, and the mechanical transmission. Suppose the efficiencies of the motor and power electronics are 0.85 and 0.95, respectively. Then, the peak power of the battery must be around 34.7 kW for this example. Table 2 lists the motor power and the battery power in FTP75 urban, highway, LA92 and US06 driving cycles.

**Plug-in Hybrid Electric Vehicles. Table 1** Vehicle parameters used in power computation

Vehicle mass (kg)	1,700
Rolling resistance coefficient	0.01
Aerodynamic drag coefficient	0.3
Front area( $\text{m}^2$ )	2.2
Rotational inertia factor	1.05



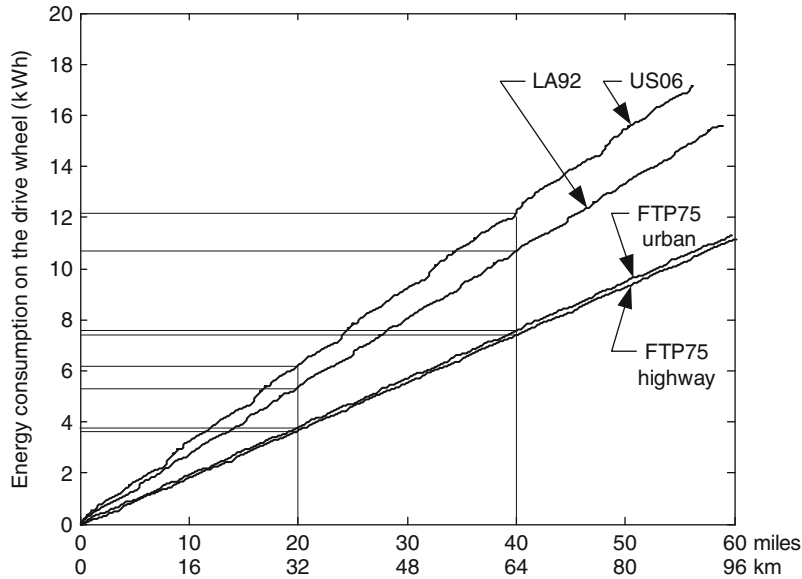
**Plug-in Hybrid Electric Vehicles. Figure 2**

Vehicle speed and traction power in FPT 75 urban driving cycle



Plug-in Hybrid Electric Vehicles. Table 2 Powers of motor and energy storage in typical driving cycles

	FTP 75 urban	FTP 75 highway	LA92	US06
Motorpower@vehicle speed	28 @50 km/h (31 mph)	32 @72 km/h (57 km/h)	55@57 km/h (36 mph)	98 @117 km/h (73 mph)
Energy storage power	35.7	39	68.5	121



Plug-in Hybrid Electric Vehicles. Figure 3

Energy consumption on the drive wheels versus driving distance in typical driving cycles

### Energy Consumption in Typical Driving Cycles

The all-electric range (AER) of a PHEV is determined by the energy capacity of the battery and the energy consumption in drive cycle. The amount of energy consumed in a typical drive cycle can be obtained by integrating (1) over the driving time period, as shown in Fig. 3, in which no regenerative braking is included. Considering the energy losses in the power electronics, motor, and transmission, the usable energy in the battery for 20 and 40 miles (32 and 64 km) of AER in typical drive cycles is listed in Table 3.

In vehicle design, a proper reference drive cycle should be selected. An aggressive drive cycle, such as US06, needs a large motor drive and battery which also leads to good vehicle acceleration and gradeability performance. On the other hand, a mild drive cycle, such as FTP75, needs a small motor drive and battery, but also leads to a sluggish vehicle performance.

Plug-in Hybrid Electric Vehicles. Table 3 Energy consumption in typical driving cycles

	FTP 75 Urban	FTP 75 highway	LA92	US 06
20 miles (32 km)	5.2	5.14	7.29	8.4
40 miles (64 km)	10.4	10.28	14.58	16.8

### Operation Strategies

A PHEV may operate in an all-electric range (AER) mode or a blended EV/HEV operation strategy.

#### AER Mode Operation Strategy

The principle of this operation strategy is to use the energy of the battery exclusively. One possibility is to allow the driver to manually select between the CS HEV

mode and the pure EV mode. The availability of AER with sufficient range allows the vehicle to be driven in areas where emissions are restricted. This strategy provides flexibility for the driver to choose the time when the pure EV mode is used. For example, in a trip that includes a distance where pure EV operation is required, the driver can select the pure EV mode just prior to entering this area in order to have sufficient range. In other places, the vehicle may be operated in pure EV mode or CS HEV mode, depending on the charge status in the battery and the power demand. In normal conditions where the trip does not have an imperative pure EV operation, the driver may select pure EV mode at start of the trip in order to fully use the energy of the battery to save petroleum fuel, until the charge in the battery reaches its design specified level at which the CS HEV mode will start automatically.

The energy drawn from the battery in the pure EV mode can be expressed as

$$E_d = \int_0^T P_d dt, \quad (2)$$

where  $P_d$  is the discharging power of the battery,  $T$  is the total driving time,  $P_d$  is the traction power. During braking, regenerative braking may be used and part of the braking energy is recovered and restored to the battery. The stored energy can be expressed as:

$$E_c = \int_0^T P_c dt, \quad (3)$$

where  $P_c$  is the charging power of energy storage during braking. It should be noted that regenerative braking power is smaller than the total braking power of the vehicle and only part of braking energy can be recovered [7–9]. The SOC of the battery can be expressed as

$$\text{SOC} = \text{SOC}_0 - \frac{E_d - E_c}{E_t}, \quad (4)$$

where  $\text{SOC}_0$  is the initial value of the SOC that may be equal to 1, and  $E_t$  is the total energy of the energy storage with  $\text{SOC} = 1$ . When the SOC of the battery drops to a specified value (0.3 for example, depending on the operation characteristics of the battery), the CS mode starts.

In the CS mode, the SOC of the battery should be maintained around the specified value. Here, an engine

control constrained to on/off control strategy is used, as described below.

1. When the acceleration pedal traction power command is greater than the engine maximum power (full open throttle), the engine is operated with full open throttle and the electric motor supplies the rest of the power to the drive train. That is,

$$P_{eng} = P_{eng-max}, \quad (5)$$

and

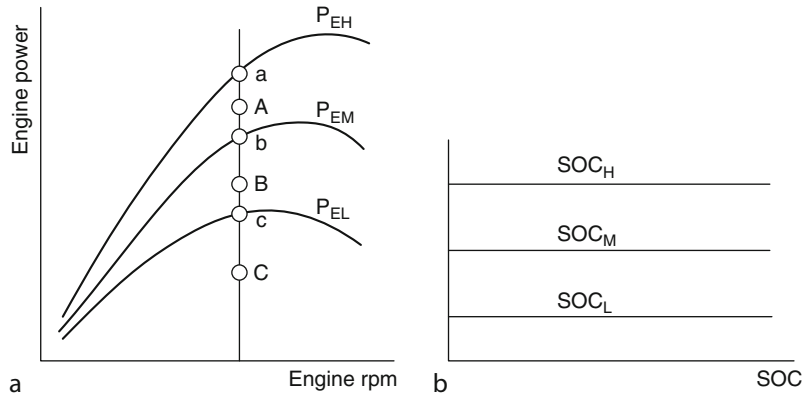
$$P_{mot} = P_{tra} - P_{eng}, \quad (6)$$

where  $P_{eng}$  is the engine power,  $P_{eng-max}$  is the engine maximum power with full open throttle,  $P_{mot}$  is the motor traction power, and  $P_{tra}$  is the commanded traction power of the driver through the acceleration pedal.

2. When the commanded traction power is smaller than the engine maximum power with full open throttle, the engine and motor operation depends on the SOC of the battery, as described below and depicted in Fig. 4.

(a) When the commanded traction power, represented by point *A* in Fig.4a, falls in the engine power range of high to medium threshold (labeled by  $P_{EH}$  and  $P_{EM}$  in Fig. 4a), and the battery SOC is lower than  $\text{SOC}_L$ , the engine is operated at point *a* (maximum engine power) so that it has excess power to charge the energy storage. Otherwise if SOC is greater than  $\text{SOC}_L$ , the engine is operated to produce power that is equal to the commanded power (point *A*).

(b) When the commanded traction power, represented by point *B* in Fig.4a, falls in the engine power range of medium to low (between  $P_{EM}$  and  $P_{EL}$ ) as shown in Fig. 4a, and if battery SOC is lower than  $\text{SOC}_L$ , the engine is operated at point *a* with full open throttle so as to quickly raise the battery SOC with a large charging power ( $P_{Ea}-P_B$ , where  $P_{Ea}$  is the engine power at point *a* and  $P_B$  is traction power at point *B*). Otherwise if the battery SOC is higher than  $\text{SOC}_L$  but lower than  $\text{SOC}_M$ , the engine is operated at point *b* and the relatively small engine power is used to charge the battery ( $P_{Eb}-P_B$ ). However, if the battery SOC is higher than  $\text{SOC}_M$ , engine alone



Plug-in Hybrid Electric Vehicles. Figure 4

Constraint engine on/off control strategy, (a) engine operating regions, (b) energy storage

traction is used, e.g., the engine is controlled to produce power just equal to the commanded traction power and the battery is neither charged nor discharged.

- (c) When the commanded traction power, represented by point C in Fig.4a, falls in the low power range (below  $P_{EL}$  as shown in Fig. 4a) and if the battery SOC is below  $SOC_L$ , the engine is operated at point b to produce large battery charging power ( $P_{Eb}-P_C$ ). Whereas if the battery SOC is in a range larger than  $SOC_L$ , but lower than  $SOC_M$ , the engine is operated at point c, and the charging power of energy storage is  $P_{Ec}-P_C$ . However, if the battery SOC is higher than  $SOC_M$ , the engine is shut down to avoid low engine operating efficiency, and the vehicle is operated with pure EV mode.
3. When braking is applied, the electric motor is always operated in regenerative braking [7–9].

The thresholds of the engine power ( $P_{EH}$ ,  $P_{EM}$ , and  $P_{EL}$  as shown in Fig.4a) are related to the engine fuel consumption characteristics, and should be set such that the battery SOC is maintained in a region lower than  $SOC_H$  and higher than  $SOC_L$ . The thresholds of battery SOC ( $SOC_H$ ,  $SOC_M$ , and  $SOC_L$ ) should be set such that it has sufficient power to support motor operation.

The control strategy discussed above is only one of many possible options. Other control strategies may be used.

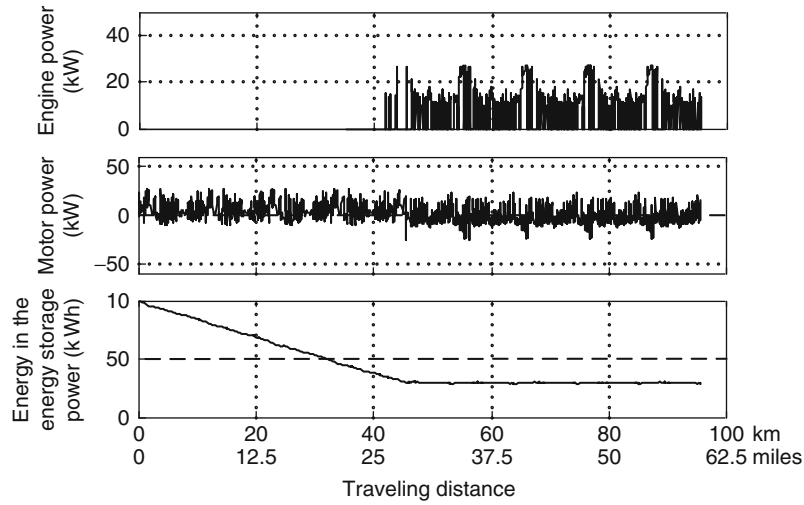
A PHEV which has the parameters listed in Table 1 has been simulated by using simulation software

developed by The Advanced Vehicle Systems Research Program at Texas A&M University. In this simulation, a 400 W electric load for the vehicle accessories is added. The total energy in the fully charged battery is 10 kWh. The simulation sequentially ran nine cycles of FTP75 urban drive cycle, and the pure EV mode was started at the beginning of the simulation until the SOC reached about 30%, beyond which, charge-sustained mode was started.

The simulation results in FTP75 urban drive cycle are shown in Figs. 5, 6, 7. The all-electric range is around 42 km in which about 7 kWh of electric energy is consumed. The engine operating points are placed within the favorite region of the engine fuel consumption map as shown in Fig. 6. The fuel and electric energy consumption are shown in Fig. 7. It can be seen that when the traveling distance is less than four cycles (42 km or 26 miles), the vehicle operates in pure EV mode and completely displaces the petroleum fuel with electricity. The total electric energy consumed is about 7.1 and 15.5 kWh per 100 km, or 4.05 miles/kWh. With the increase in the total traveling distance, the percentage of the fuel displacement decreases since the charge-sustained operation takes a larger percentages of the travel distance. For nine sequential drive cycle (96 km or 60 miles), the fuel and electrical energy consumptions are about 3.2 L/100 km or 74 mpg, and 7.42 kWh/100 km or 8.43 miles/kWh.

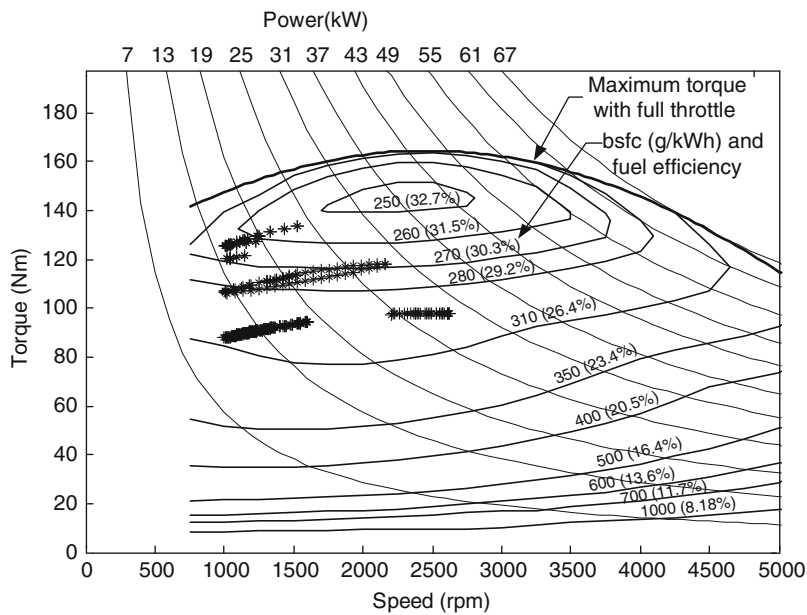
### Blended Control Strategy

Unlike the AER control strategy, the blended control strategy uses both engine and motor for traction with



Plug-in Hybrid Electric Vehicles. Figure 5

Profile of engine power, motor power, and energy in the energy storage along with traveling time in FTP75 urban driving cycle



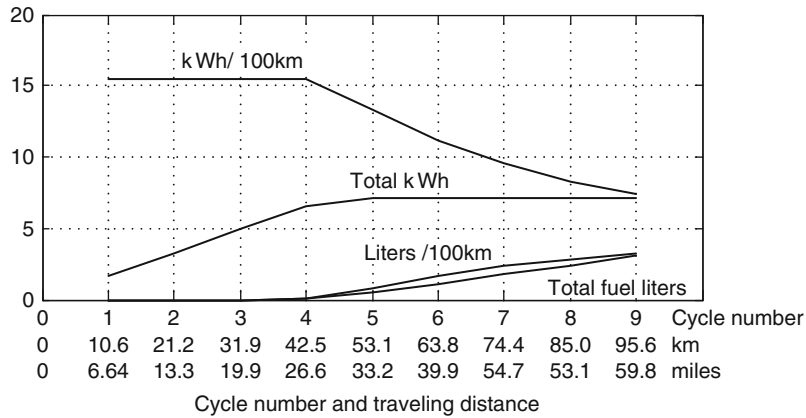
Plug-in Hybrid Electric Vehicles. Figure 6

Engine operating points on the fuel consumption map in FTP75 urban drive cycle

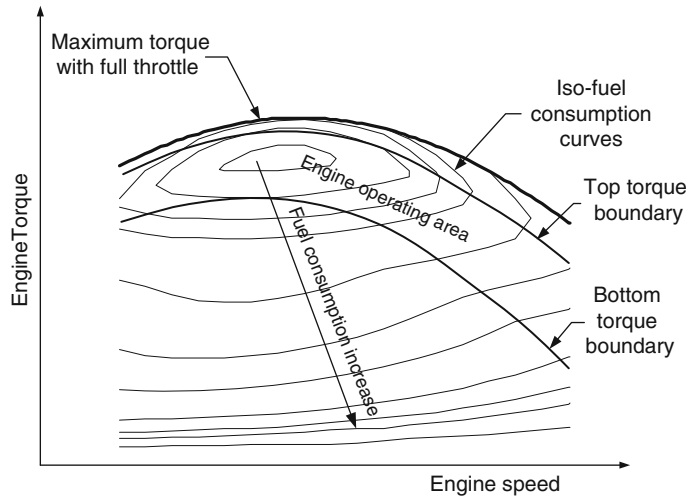
charge-depleting (CD) mode until the battery SOC reaches a specified low level, beyond which the vehicle operates in the charge-sustained (CS) mode.

In the CD mode, both the engine and the motor may operate at the same time. The range before

entering CS mode is longer than that with AER control strategy. New control strategies are needed to control the engine and motor to meet the load demand. There are many possible control strategies. The following is one in which the engine and motor alternatively propel



**Plug-in Hybrid Electric Vehicles. Figure 7**  
 Fuel and electric energy consumption in FTP75 urban drive cycle



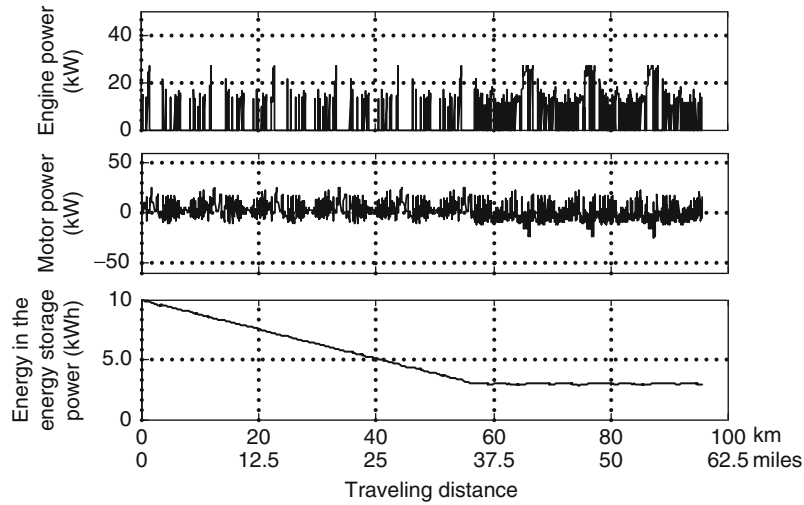
**Plug-in Hybrid Electric Vehicles. Figure 8**  
 The operation area of the engine in the CD mode

the vehicle with no battery charging from the engine. The engine is constrained to operate in its optimal fuel economy region. The details are described as follows.

Figure 8 schematically shows the engine operating area. When the requested engine torque is larger than the top torque boundary, the engine is controlled to operate on this boundary and the remaining torque is supplied by the electric motor. When the requested engine torque falls in the area between the top and bottom boundaries, the engine alone propels the vehicle. When the requested engine torque is below the

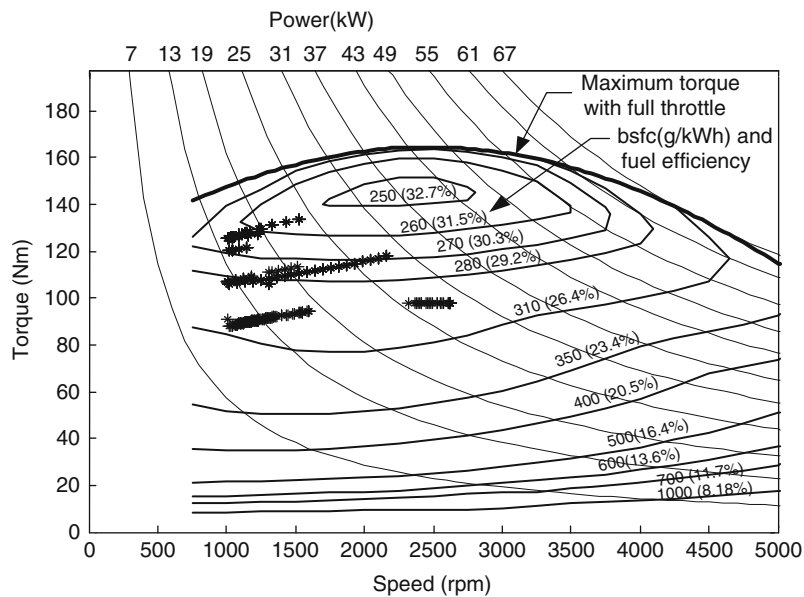
bottom torque boundary, the engine is shut down and the electric motor alone propels the vehicle. In this way, the engine operation is constrained to within its optimal fuel economy region. Since there is no battery charging from the engine, the energy level in the battery continuously drops toward its specified low level. Then the vehicle goes into CS mode, in which the constrained engine on/off control strategy, or some other control strategy, is used as discussed above.

The example vehicle mentioned above has been simulated with the control strategy discussed above in



Plug-in Hybrid Electric Vehicles. Figure 9

Profile of engine power, motor power, and energy in the energy storage along with traveling time in FTP75 urban drive cycle



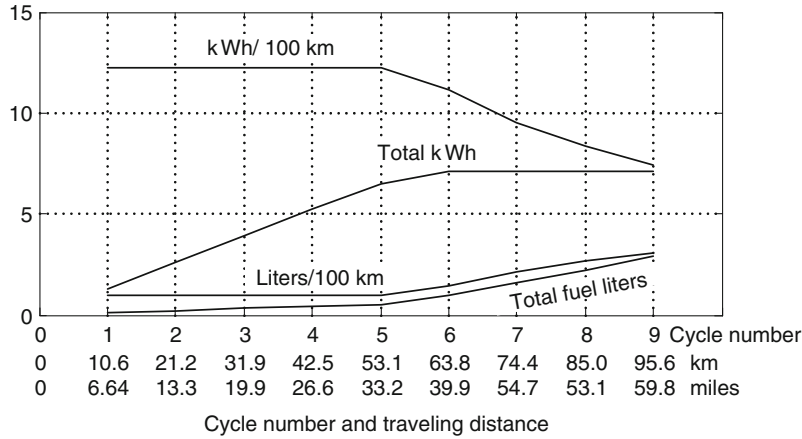
Plug-in Hybrid Electric Vehicles. Figure 10

Engine operating points on the fuel consumption map in FTP75 urban drive cycle

nine sequences of FTP 75 urban driving cycle. The results are presented in Figs. 9, 10, 11. Figure 9 shows that the battery energy continuously drops to the specified value during CD mode, at which point the engine joins the effort traction occasionally, operating in its optimal region as shown in Fig. 8 but does not charge

the battery. Figure 9 shows that the CD range is longer than the all-electric range as shown in Fig. 5.

As indicated in Fig. 10, the engine operating points overlap the favored region of the engine fuel consumption map. Figure 11 indicates that the fuel consumption in the CD range (0–53 km) is around



**Plug-in Hybrid Electric Vehicles. Figure 11**  
 Fuel and electric energy consumption in FTP75 urban drive cycle

1 L/100 km (235 mpg) and electric energy consumption is around 12.5 kWh/100 km (5 mile/kWh). However, with the increase in the travel distance, the fuel liters per 100 km increases and kWh per 100 km decreases since the ratio of the charge-sustained operation distance to the total trip distance increases.

**Battery Design**

The amount of energy in the battery determines the all-electric range (AER) or the charge-depletion range (CDR). It is also closely related to fuel consumption, fuel displacement, initial cost, and operating costs. Through simulation, similar to those performed above, the usable energy in the battery can be determined. The total energy capacity can be obtained from

$$E_c = \frac{E_{usable}}{SOC_{top} - SOC_{bottom}}, \tag{7}$$

where the  $E_{usable}$  is the usable energy in the battery consumed in the AER or CDR modes,  $SOC_{top}$  is the top SOC with fully charged battery, which usually equals 1, and  $SOC_{bottom}$  is the SOC of the battery at which the operation mode is switched from AER or CDR modes to CS mode. In the example above, the usable energy is about 7 Wh (refer to Figs. 5 and 9). Suppose that the SOC operating window is 0.7 (from 1 to 0.3). Then, the total energy capacity of the battery is about 10 kWh.

It should be noted that the depth of discharge (DoD) of batteries is closely related to battery life.

Figure 12 illustrates the cycle life for NiMH and Li-ion batteries [5]. If one deep discharge per day is supposed, a total 4,000+ deep charges would be required for a 10–15 year lifetime. With the characteristics shown in Fig. 12, 70% depth of discharge for NiMH and 50% for Li-ion batteries may be the proper designs.

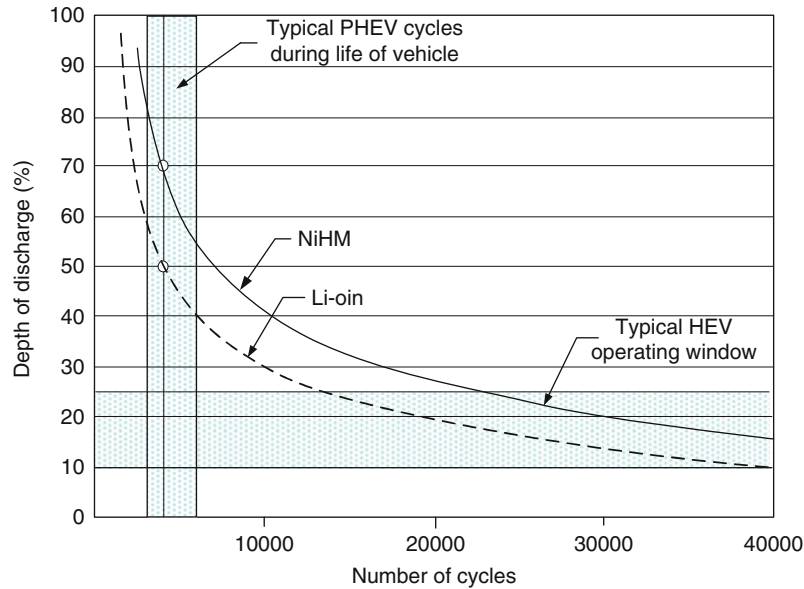
In the battery design, it should be ensured that the energy storage can supply sufficient power to support the vehicle when its SOC is at a low level (0.3 for NiMH battery and 0.5 for Li-ion battery, for example, as indicated in Fig. 12).

Energy/power ratio of a battery is a good measure of its suitability. The size of the battery will be minimized when its energy/power ratio equals that required. The energy/power ratio is defined as:

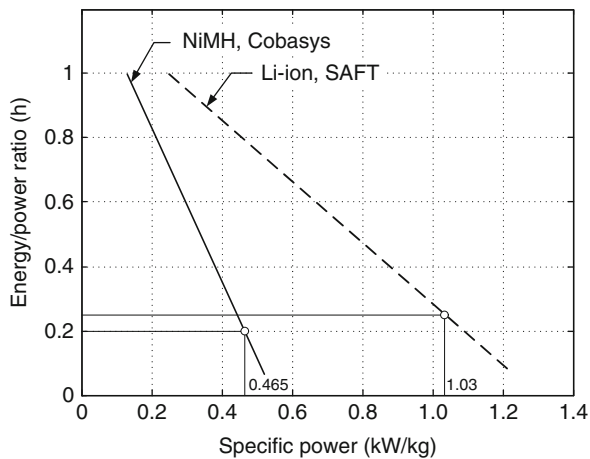
$$R_{e/p} = \frac{Total\ Energy}{Power@operating\ SOC}, \tag{8}$$

In the example vehicle simulated above, the total energy required is around 10 kWh for NiMH battery (0.7DoD) and 14 kWh for Li-ion battery (0.5 DoD). The power required is about 60 kW (refer to Table 2 for LA92 drive cycle) which is defined at 30% of SOC for NiMH battery and 50% of SOC for Li-ion battery. The required energy/power ratio is 0.167 h for NiMH battery at 30% of SOC and 0.233 h for Li-ion battery at 50% of SOC.

Figure 13 shows the energy/power ratio versus the specific power of NiMH and Li-ion batteries [5].



**Plug-in Hybrid Electric Vehicles. Figure 12**  
Cycle life characteristics of Varta battery technologies [5]



**Plug-in Hybrid Electric Vehicles. Figure 13**  
Typical energy/power ratio versus specific power

A 0.2 h of energy/power ratio (slightly larger than 0.166 h for safety) for Cobasys NiMH battery yields a total weight of 129 kg ( $60/0.465$ ), which carries 12 kWh of total energy ( $0.2 \times 60$ ). However, 0.25 h (slightly larger when 0.233 for safety) of SAFT Li-ion battery yields a total weight of 58 kg ( $60/1.03$ ) which carries 15 kWh of energy. Obviously, the SAFT Li-ion battery is superior to the Cobasys NiMH battery.

However, other factors have to be considered such as, cost, safety, etc.

### Summary and Future Directions

Plug-in hybrid electric vehicle (PHEV) can displace a significant amount of petroleum fuel with electric energy. This technology can significantly change the makeup of transportation fuel supply. Design methodologies presented in this entry can be used to design a drive train that has a specified all-electric range or charge-depletion range. The control strategies developed here can be used in vehicle control to realize all-electric operation, charge depletion, and charge-sustained operations. These control strategies can also operate the engine always within its low fuel consumption region, thus, yielding high overall efficiency. These design methodology and control strategies have been validated by simulation of passenger car driving in FTP 75, a typical urban drive cycle.

Plug-in hybrid vehicles are now being introduced by manufacturers, such as the Chevrolet Volt by GM. It is anticipated that PHEVs will occupy a larger segment of the automotive market in the coming years. These will take various forms and design philosophies to fit the market demands, price points, and fuel and environmental requirements.



## Bibliography

1. Ehsani M, Gao Y, Gays SE, Emadi A (2005) Modern electric, hybrid electric and fuel cell vehicle – Fundamentals, theory and design. CRC Press
2. Simpson A (2006) Cost-benefit analysis of plug-in hybrid electric vehicle technology. Presented at the 22nd international battery, hybrid and fuel cell electric vehicle symposium and exhibition (EVS-22), Yokohama, 23–28 Oct 2006
3. Gonder J, Markel T (2007) Energy management strategies for plug-in hybrid electric vehicles. SAE paper # 2007-01-0290, Detroit
4. Markel T, Simpson A (2005) Energy storage systems considerations for grid-charged hybrid electric vehicles. Presented in vehicle power and propulsion, 2005 IEEE conference, Chicago, 7–9 Sep 2005
5. Markel T, Simpson A (2006) Plug-in hybrid electric vehicle energy storage system design. Presented at advanced automotive battery conference, Baltimore, 17–19 May 2006
6. Markel T, Wipke K (2001) Modeling grid-connected hybrid electric vehicles using ADVISOR. Presented in the, The 16th annual battery conference on application and advances, Long Beach, California
7. Gao Y, Ehsani M (2001) Electronic braking system of EV and HEV – Integration of regenerative braking, automatic braking force control and ABS. SAE J Passenger Cars Electron Electr Syst 110. Paper No. 2001-01-2478
8. Gao Y, Chen L, Ehsani M (1999) Investigation of the effectiveness of regenerative braking for EV and HEV. SAE J Passenger Cars 108. Paper No. 1999-01-2901
9. Gao Y, Chu L, Ehsani M (2007) Design and control principle of hybrid braking system for EV, HEV and FCV. In: IEEE on vehicle power and propulsion conference, Arlington, 9–12 Sep 2007

## Polio and Its Epidemiology

LESTER M. SHULMAN

Central Virology Laboratory, Laboratory of Environmental Virology at Sheba Medical Center, Public Health Services, Israel Ministry of Health, Department of Epidemiology and Preventive Medicine, School of Public Health, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel

### Article Outline

Glossary

A Brief Definition of Polio and Its Importance

Introduction

The Epidemiology of Polio

Future Directions: The Endgame Stage of Eradication and Sustainability of Postpolio Eradication  
Bibliography

### Glossary

**ACPE** The Advisory Committee on Polio Eradication

**AFP** Acute flaccid paralysis.

**AFP surveillance** Characterization of enteroviruses in stool samples from all AFP cases especially in individuals under 15 years of age to rule-in or rule-out etiology by polioviruses.

**aVDPV** A vaccine-derived poliovirus isolate whose evolutionary path is unknown or ambiguous.

**bOPV** Bivalent oral polio vaccine (usually containing serotypes 1 and 3).

**BSL** Biosafety standard level.

**Capsid** The protein shell that surrounds a virus particle.

**Capsomere** One of the individual morphological units that make up the viral capsid.

**CDC** US Centers for Disease Control and Prevention

**CD155 or PVr** The human encoded cell receptor for poliovirus, a member of the immunoglobulin superfamily.

**Codon** A sequence of three adjacent nucleotides on a strand of DNA or RNA that specifies which specific amino acid will be incorporated into a protein.

**Codon bias** Unequal usage of synonymous codons (different codons that specify the same amino acid)

**CPE** Cytopathic effect.

**cVDPV** A circulating vaccine-derived poliovirus, that is, a poliovirus that has evolved from vaccine during person-to-person transmission.

**eIPV** Enhanced inactivated polio vaccine.

**Emergence** The appearance of a pathogen in a previously pathogen-free area.

**Endemic** The constant presence of a disease to a greater or lesser extent in a particular locality.

**Enteroviruses** Any of >80 different species of polioviruses, coxsackie viruses, echoviruses, and enteroviruses belonging to the genus *Enterovirus* in the family *Picornaviridae*.

**Environmental surveillance (as related to polioviruses)** Investigation of sewage and recreational water for the presence of poliovirus as an

- indication of the presence of poliovirus-infected individuals in a community.
- EPI** Expanded Program on Immunization.
- Epidemic** A rapid spread of disease into a disease-free area or spread of a disease to more than the usual number of persons affected in a region with disease.
- Epitopes** The component of an antigen that is recognized by and binds to an antibody.
- Eradication** The complete elimination of all incidence of disease and/or the presence of the agent that causes the disease.
- Evolution** Change in the genetic composition of a population or the genome of a given organism during successive generations.
- GAP** Global action plan for laboratory containment.
- GAP I** GAP phase I – plan for identifying all known and potential sources of poliovirus especially wild polioviruses within each country.
- GAP II** GAP phase II – plan for laboratory containment of wild polioviruses.
- GAP III** GAP phase III – plan to minimize post-eradication poliovirus facility-associated risks.
- GAVI** Global alliance for vaccines and immunization
- Genetic recombination (of polioviruses)** A situation where one portion of the genome of a poliovirus is replaced through a covalent linkage with the equivalent segment from another poliovirus or non-polio enterovirus.
- Genotype** The genetic makeup of an organism as distinguished from its physical characteristics.
- GMT** Geometric mean titer, usually calculated according to Karber.
- GOARN** Global Outbreak Alert and Response Network.
- GPEI or GEI** The Global Poliomyelitis Eradication Initiative of the WHO, adopted in 1988.
- GPLN** Global Polio Laboratory Network.
- Hydrophobic pocket** A hydrophobic space located under the binding site for the host encoded viral receptor that is located on the bottom of the canyon surrounding the fivefold axis of symmetry of the enteroviral capsid.
- Hydrophobic pocket factors** Small hydrophobic molecules that occupy the hydrophobic pocket and that may regulate the host receptor viral capsid interaction.
- Immunodeficient** Lacking one of the components of the immune system.
- Immunogenicity** The relative ability of a molecule to elicit an immune response.
- i.d.** Intradermal or under the skin.
- i.m.** Intramuscular.
- i.n.** Intranasal.
- i.p.** Intraperitoneal.
- i.t.** Intrathecal.
- Infection** Establishment and growth of an infectious agent in the body.
- IPV** Inactivated poliovirus vaccine.
- IRES** Internal ribosome entry site.
- ITD** Intratypic differentiation (determination if a virus isolate is vaccine, vaccine-derived, or wild).
- iVDPV** A vaccine-derived poliovirus that has diverged from its respective oral poliovirus serotype during persistent infection of an immunodeficient host.
- Lineages** A group of organisms that are closely related genetically.
- MAPREC** Mutant analysis by PCR and restriction fragment enzyme cleavage to measure reversion of attenuation sites in vaccine strains.
- MNVT** Monkey neurovirulence test, an *in vivo* neurovirulence test in monkeys.
- mOPV** Monovalent OPV.
- Neurovirulence** The ability of the poliovirus to infect and damage nerve cells causing disease of the nervous system.
- Neurovirulence attenuation sites** Specific nucleotide positions along the poliovirus genome where the specific nucleotide present at that site will influence whether or not an individual polioviral isolate will be neurovirulent.
- Neutralizing antigenic sites** Epitopes of the poliovirus that induce neutralizing antibodies.
- NID** National immunization day.
- NSL** Non-Sabin-like (wild) virus of vaccine-derived poliovirus (based on results of certain ITD tests).
- Major disease** Poliomyelitis, AFP, or cases of infection with polio that involves invasion and permanent damage to the nervous system.
- Microarray** A technology used to study many genes at once using thousands of different short molecular sequences at known position on solid support to hybridize to complementary nucleic acid sequences from different sources.
- Minor disease** Nonspecific illness caused by poliovirus that may include upper respiratory tract

symptoms (sore throat and fever), gastroenteritis (nausea vomiting, abdominal pain, constipation or diarrhea), and influenza-like illness.

**NCCs** National Certification Committees.

**NGOs** Nongovernmental organizations.

**NIDs** National immunization days.

**NPEV** Non-polio enteroviruses.

**Nonstructural genes** Viral genes encoding proteins that are not incorporated into the structure of the capsid.

**Oligonucleotide** A short sequence of nucleotides frequently synthetic.

**OPV** Live attenuated oral poliovirus vaccine.

**Outbreaks** (Under eradication conditions) even the presence of a single case of paralytic poliomyelitis.

**Persistent poliovirus infection** An infection associated with an immunodeficient host where virus is not cleared but continues to replicate for an indefinite period of time.

**Phylogenetic tree** A diagram with branches showing the inferred evolutionary relationships among various biological entities.

**Picornaviridae** A viral family made up of the small (18–30 nm) ether-sensitive single stranded, positive-sense RNA viruses that lack an envelope.

**Poliomyelitis** The infectious disease caused by poliovirus involving inflammation of motor neurons of the spinal cord and brainstem that leads to acute paralysis followed by atrophy of the muscles innervated by the infected motor neurons.

**Poliovirus** One of three serotypes of picornaviruses that can cause acute flaccid paralysis and whose cell receptor is CD155.

**Polypeptide** A small molecule constructed from linked amino acids.

**Postpolio syndrome** Slow progressive muscle pain and weakness that reappears 30 or 40 years after paralysis caused by a poliovirus infection affecting muscles previously affected by polio as well as muscles that may not have been affected.

**Posttranslational processing** Any modification of a protein after it has been translated.

**Provoked poliomyelitis** Poliomyelitis resulting from physical trauma during infection with poliovirus.

**Proofreading** An enzymatic process that checks whether a newly incorporated nucleotide in a nascent chain is the correct complement of its corresponding nucleotide in the template.

**PVR** The poliovirus receptor, CD155.

**PVR Tg21 transgenic mouse** A mouse that has been genetically modified to express the human poliovirus receptor.

**Quasispecies** A term used to describe a cluster, cloud, or swarm of viruses with minor differences in nucleotide sequence that arise during replication as a consequence of polymerase incorporation errors.

**Rearrangement (in relation to polioviruses)** A structural alteration in the genomic sequence occurring during coinfection with two or more viruses resulting in a new genome in which parts are from different parental polio or non-polio enteroviruses

**Reemergence** Emergence after an absence.

**RCC** Regional Certification Committees.

**RCT** Reproductive capacity temperature, the temperature at which viruses can replicate.

**RNA-dependent RNA polymerase** A viral encoded polymerase that synthesizes a complementary RNA strand from an RNA template.

**RRL** Regional Reference Laboratory of the Global Polio Laboratory Network.

**SAGE** Strategic Advisory Group of Experts on Immunization.

**Serotype** A group of closely related virus expressing a common set of antigens.

**Seroconversion** Appearance of antibodies following exposure to antigen in seronegative person, or  $\geq 4$ -fold increase in titer of previously immune person.

**Seroconversion index** The mean seroconversion rate against all three poliovirus serotypes.

**SL** Sabin-like poliovirus (based on result from some ITD tests).

**SIA** Supplemental immunization activities (such as NIDs, SNIDS, and mop-ups).

**Silent circulation** Person-to-person transmission of virus in a community in the absence of cases of AFP.

**Silent infection** Asymptomatic infection.

**Silent presence** The presence of a virus in the absence of clinical cases and the absence of person-to-person transmission.

**SNIDS** Sub-national Immunization Days.

**Stakeholders** All governmental and nongovernmental agencies involved in the GPEI.

**Structural genes (in relation to polio)** Genes encoding viral capsid proteins.

**Synonymous nucleotide substitutions** Changes in the nucleotide sequence that do not result in a change in encoded amino acid.

**TAG** Technical Advisory Group.

**TD** Typic differentiation (determination of the serotype of a poliovirus isolate).

**TOPV** Trivalent oral polio vaccine containing all three serotypes of attenuated poliovirus.

**Transition** The substitution of a purine nucleotide with the other purine, or a pyrimidine nucleotide with the other pyrimidine.

**Transversions** The substitution of a pyrimidine nucleotide by a purine nucleotide or vice versa.

**UNICEF** United Nations Children's Fund.

**Vaccine strains** Poliovirus strains approved by the WHO for production of live and inactivated polio vaccines.

**VAPP** Vaccine-associated paralytic poliomyelitis.

**VDPV** A vaccine-derived poliovirus that has diverged through evolution from its respective live poliovirus vaccine strain serotype by more than 1% (serotypes 1 and 3) or more than 0.6% (serotype 2) of its respective VP1 capsid protein.

**Viremia** The presence of virus in the bloodstream during an infection.

**VPg** Viral protein genome linked – 22 amino acid protein covalently linked to genome and complementary negative strand.

**VP1** Viral capsid protein 1.

**VP2** Viral capsid protein 2.

**VP3** Viral capsid protein 3.

**VP4** Viral capsid protein 4.

**WHA** World Health Assembly.

**WHO** World Health Organization.

**WPV or wild poliovirus** Any poliovirus that is not derived from attenuated oral polio vaccine strains

**3'UTR** The untranslated region of the polioviral genome that is located 3' of the open reading frame that encodes the viral polyprotein.

**3D<sup>pol</sup>** Viral encoded RNA-dependent RNA polymerase.

**5'UTR** A highly structured, untranslated area of the polioviral genome located 5' to the open reading frame that encodes the viral polyprotein. The 5'UTR is covalently linked on its 5' base to viral protein VPg.

## A Brief Definition of Polio and Its Importance

The word “polio” has been used to describe both a disease and the disease agent. Among current methods to measure the importance of or interest in a topic is to run a general web search for the term and to search the scientific literature in PubMed. A Google web search of the word “polio” in Aug 2010 yielded 31,100,000 hits, while a search in PubMed yielded 22,000 articles and 826 review articles. This review will concentrate on those aspects of the epidemiology of polio as it relates to disease eradication and the sustainability of this effort. The terms “polio” and “poliomyelitis” will be used when describing the disease and “poliovirus” and related terms such as “polio vaccine” will be used to describe the agent that causes the disease.

In order to understand the epidemiology of polio, it is important to understand the adversary. Toward this goal, this chapter starts with a detailed physical characterization of polioviruses and the pathological effects caused by poliovirus infections that are most relevant to understanding the epidemiology of polio. This is followed by a description of the global efforts to eradicate poliomyelitis and the viral agent causing the disease, and concludes with a discussion of the future directions needed to achieve and sustain eradication and prevent reemergence. Smallpox was the first human disease to be eradicated and we are currently in the endgame of eradication of polio as the second. Polio eradication is currently the largest public health program in the world and has involved both health professionals and more than ten million volunteers in all countries since the inception of the Global Poliomyelitis Eradication Initiative by the World Health Assembly in 1988 [1].

## Introduction

The road toward polio eradication has been long [2] and by no means smooth. Important milestones along the march toward recognition and understanding the disease, identification of its causative agent, and toward prevention and eradication will be briefly discussed in the introduction (see also Fig. 1). A poliovirus isolate is classified as vaccine, vaccine-derived (VDPV), or wild-type poliovirus based on the percent nucleotide sequence homology between its capsid protein VP1

Major milestones along the road towards polio eradication.

- 1400 BCE First pictorial record of a person with poliomyelitis.
- 1789 First modern characterization of polio as a "debility of the lower extremities" by Underwood.
- 1813 More complete detailed description of polio by Monteggia.
- 1855 Description of "infantile paralysis by Heine based on systematic investigation of cases started in the 1840's.
- 1863 Confirmation by biopsy for motor neuron involvement in polio by Cornil.
- 1870 Detailed description of physiological changes in the anterior horn of the spinal cord by Charcot and Joffroy.
- 1889 Medin realizes that paralytic cases occur in only a small number of infected individuals during epidemics.
- Late 19<sup>th</sup> Century The emergence of outbreaks of poliomyelitis.
- 1905 Recognition of the infectious nature of polio by Wickman confirming Medin's earlier observations.
- 1908-9 The discovery of poliovirus as the causative agent of poliomyelitis by Landsteiner and Popper.
- 1920'2 Establishment of the first national rehabilitation center.
- 1929 Drinker develops the iron lung.
- 1931 Burnet and MacNamara report more than one non-cross reacting antigenic strain of polio; culminates in the conclusion in 1951 that there were only three serotypes.
- 1935-36 First clinical trials with inactivated polio vaccine by Brodie and Park and with live attenuated vaccine by Kolmer.
- 1937 Establishment of the first NGOs to fund support of polio victims and research: The National Foundation of Infantile Paralysis and The March of Dimes.
- 1939 Armstrong was the first to grow poliovirus in a non-primate (rodent) host.
- 1940's Kinny introduces the concept of supportive rehabilitation.
- 1949 First in vitro passages in tissue cultures by Enders, Weller, and Robbins
- 1950's Development and testing of attenuated vaccines by Kaprowski (1950), Sabin (1956-57) and Cox (1958).
- 1950's Bottiger and Kaprowski observed that live vaccine spreads to contacts.
- 1951 Conclusion that there were only three serotypes of poliovirus.
- 1951-61 Immunization of > 11 million children with Sabin and with Kaprowski live oral vaccine strains.
- 1953-54 Salk develops and tests inactivated poliovirus vaccine licensed for use by 1955.
- 1954 First use of live vaccine by Kaprowski to control a large outbreak.
- 1954 Development of standard plaque assays for quantification of virus by Dulbecco and Voigt.
- 1955 The Cutter Incident in which 400,000 children were immunized with inadequately inactivated wild poliovirus.
- 1962 Detection of the first persistent infections with vaccine-derived polioviruses.
- 1968 Introduction of microcarrier cell systems for uniform large-scale vaccine production by van Wezel followed by the use of pathogen cell-free cell systems in the early 1990's.
- 1979 The development of primate model to test neurovirulence of poliovirus strains.
- 1979 Last case of wild polio in the Unites States.
- 1985 Pan Americans Health Organization and CDC establish the Latin American Regional Polio Network.
  
- 1986 Development of murine L20B cells transformed with, and expressing the human receptor allowing selective growth of poliovirus while non-permissive for most other human enteroviruses.

### Polio and Its Epidemiology. Figure 1

(Continued)

and that of the corresponding OPV vaccine serotype. An isolate with VP1 homology of 99–100% is classified as vaccine virus, 85–99% as VDPV, and > 85% as wild-

type poliovirus [3]. This rule of thumb for classifying polioviruses as VDPVs has recently been modified for serotype 2 to include isolates with  $\geq 6$  nucleotide

1988	WHO establishes the Global Poliomyelitis Eradication Initiative
1989	Identification and cloning of the poliovirus receptor by Mendelsohn, Wimmer and Racaniello.
1981	Recombinant nucleic acid techniques used by Racaniello and Baltimore to prepare an infectious clone of poliovirus.
1990's=>	Development of molecular and immunological assays to identify the serotype and determine the vaccine or non-vaccine origin of poliovirus isolates.
1990-91	Development of a transgenic mouse expressing the human encoded poliovirus receptor approved as a non-primate model for neurovirulence testing safety of all three serotypes of live OPV in 1999-2000.
1991	Last case of endogenous wild polio in the Western Hemisphere.
1991	Establishment of a global Polio Laboratory Network to monitor poliovirus infections throughout the world.
1998=>	Isolation and characterization of highly diverged vaccine derived polioviruses from environmental samples excreted by unknown individuals.
1999	Establishment of the Global Action Plan for laboratory containment of all wild polioviruses throughout the world.
1999	Last reported case of poliomyelitis anywhere in the world caused by a wild serotype 2 poliovirus.
2000-01	First prospective recognition that an outbreak was caused by a vaccine-derived poliovirus (in Haiti and the Dominican Republic).
2000	Last case of endogenous wild poliomyelitis in the Western Pacific Region.
2002	Last case of endogenous wild poliomyelitis in the Eastern European Region.
2002	Cello, Paul, and Wimmer synthesize infectious poliovirus from individual nucleotides.
2003	Failure to vaccinate in Nigeria leads to a large increase in the number of cases and exportation of wild polioviruses and vaccine derived -polioviruses to > 21 polio-free countries. This spread is being brought under control by local and regional vaccination campaigns using monovalent, divalent, and trivalent vaccines.
2005	Decision: successful eradication must include cessation of the routine use of OPV.
2010	Large successful clinical trials using fractional sub-dermal doses of IPV.

### Polio and Its Epidemiology. Figure 1

#### Major milestones along the road toward polio eradication

changes (i.e., <1%) and the upper limit of 15% for VP1 divergence has been eliminated (Summary of the 16th Informal Consultation on the Global Polio Laboratory Network, Geneva, Switzerland, 2010).

The earliest record attributed to polio comes from an Egyptian Stele from 1400 BCE that depicts an Egyptian high priest or official with a walking stick and withered leg that bears a striking resemblance to a recent picture of a man with poliomyelitis (Fig. 2). Polio infections from this time to the nineteenth century were endemic and usually occurred in young children where most infections were probably asymptomatic. While early descriptions of “acquired clubfoot” by Hippocrates and Galen were consistent with polio, the first modern medical characterization of polio includes descriptions of “Debility of the Lower Extremities” by Underwood in 1789, polio by Monteggia in 1813, “infantile paralysis” by Heine in 1840, and involvement of motor neurons in infantile

paralysis by Duchenne in 1855. Involvement of motor neurons was confirmed by biopsy of the brain and spinal cord of a polio victim by Cornil in 1863 and by a detailed description of physiological changes in the anterior horn of the spinal cord by Charcot and Joffroy in 1870.

A new epidemiological aspect of polio emerged in the nineteenth century, namely, the appearance of outbreaks that increasingly affected adults as well as children [4]. Paradoxically this shift from an endemic to an outbreak pattern of disease transmission may have been facilitated by a “hygiene barrier” derived from improved community sanitation that may have resulted in a shift from fecal–oral to oral–oral transmission, an increase in naïve individuals especially among older cohorts, and primary exposure of increasingly older cohorts where disease manifestation were more severe. These epidemics became more frequent by the mid-twentieth century and involved growing



### Polio and Its Epidemiology. Figure 2

*Living after paralytic poliomyelitis: then and now.* Paralytic poliomyelitis occurs after a biphasic infection where viremia in a small number of systemic infections is followed by infection of the CNS. Paralysis is a direct result of destructive replication of poliovirus in motor neurons followed by atrophy of de-energized muscles. Both pictures represent men whose skeletal muscles have been affected by infections of nerves in the anterior horn of their spinal cord. The picture on the left (a) depicts the earliest record of poliomyelitis in a man and comes from a stele from ancient Egypt created around 1500 BCE, and is strikingly similar to the image of the man in the photograph on the right (b) who has atrophy of the right foot and leg due to polio that was taken in the Far East in 2007 ((a) Egyptian Stele at the Ny Carlsberg Glyptotek Museum, Copenhagen, Denmark (GNU free documentation License). (b) Photograph #134 Centers for Disease Control and Prevention Public Image Library [CDC/NIP/Barbara Rice])

numbers of people. Wickman described the acute infectious nature of polio in his analysis of a 1905 polio outbreak in New York, confirming Medin's realization in 1889 that paralytic cases were only a small part of epidemics and that even persons with mild illness could infect others. Further complications were the observation by Burnet and MacNamara in 1931 [5] that different strains of poliovirus caused disease, but infection with some strains did not protect against subsequent infection with other strains and the observation in the 1950s that poliomyelitis could be triggered by physical injury during a poliovirus infection and that there was an increased risk of paralysis in limbs that received a mechanical stress or after

tonsillectomies [6]. Two important new concepts were the establishment of a national center for treatment of poliomyelitis victims at Warm Springs, Georgia, and the use of professional fund-raisers by President Roosevelt supported by others in the late 1920s. The non-partisan National Foundation for Infantile Paralysis and the March of Dimes established in 1937 institutionalized this fundraising effort. The iron lung, developed by Drinker in 1929, and the concept of supportive rehabilitation involving the use of hot moist packs to relieve muscle spasm and physiotherapy to maintain strength of unaffected muscle fibers promoted by Kenny in the 1940s were important advances for treatment of poliomyelitis.

The study of the pathological organism that caused poliomyelitis was enabled by the discovery of a bacteria-free “filterable” etiological agent, the poliovirus, which could pass disease from one primate to another by Landsteiner and Popper in 1909. Burnet and MacNamara realized in 1931 that there was more than one type of poliovirus since exposure to some isolates did not protect against exposure to others. By 1951, the National Foundation for Infantile Paralysis concluded that there were only three serotypes of poliovirus. The study of poliovirus was aided by (a) the first passages of poliovirus in a non-primate rodent system by Armstrong in 1939, (b) passage in tissue cultures by Enders, Weller, and Robbins in 1949 [7], (c) development of plaque assays for quantification of polio by Dulbecco and Vogt in 1954 [8], (d) the use of microcarrier cell systems for vaccine production by van Wezel in 1967 [9], (e) development of monkey neurovirulence tests in 1979 [10], (f) the use of pathogen-free diploid MRC5 cells (human fetal cells derived from normal lung tissue) and permanent cell lines like Vero (a cell line prepared from the kidney of a normal adult African green monkey) for vaccine production in the early 1990s, (g) identification and cloning of the poliovirus receptor CD155 [11], (h) development of the transgenic PVr-mouse model which expresses the human poliovirus receptor as an alternative to monkeys for neurovirulence testing [12], (i) preparation of a murine cell line, L20B, expressing the human poliovirus receptor for selective growth of poliovirus [13], and (j) the development of the immunological and molecular tools (discussed in detail below) that provide the identity the serotype of the isolate, distinguish whether its origin was from a vaccine or wild strain, and provide phylogenetic information on the evolutionary relationship to other isolates.

Advances in culturing polioviruses outlined in the previous paragraph laid the foundation for developing the vaccines that have turned polio into a vaccine-preventable disease and a candidate for eradication (see below). Early experiments and clinical trials such as those in 1935–1936 with inactivated poliovirus by Brodie and Park [14] and attenuated live vaccine by Kolmer [15] were hampered by lack of awareness until 1951 that there were three serotypes. Afterward, effective inactivated vaccine was developed and tested by Salk and coworkers starting in 1953–1954 [16, 17],

while Koprowski, Sabin, and Cox developed and tested attenuated oral vaccines in 1950 [18, 19], 1956–1957 [20], and 1958 [21], respectively. Between 1951 and 1962, 12.9 million children were vaccinated with Koprowski strains and 11 million with Sabin strains [19]. A number of important epidemiological observations were made during that time that continue to guide current vaccination strategies. For attenuated oral vaccines these included (a) the first demonstration by Koprowski of interference between poliovirus serotypes during coinfection [19], (b) a demonstration that maternal antibodies did not prevent an immune response in vaccinees under 6 months of age [19], (c) the observation by Koprowski and especially Bottiger that live vaccine spread to contacts [19], (d) a demonstration of persistence of antibodies at the same levels in vaccinated children for at least 3 years [19], (e) proof of concept by Koprowski that live polio vaccine could be effective in containing large outbreaks [19], and (f) documentation of high vaccine safety with both the Koprowski and Sabin OPV strains [4, 19, 22]. Safety issues relating to both the live and inactivated viral strains will be mentioned in discussions starting on pages 8150 and 8160. After extensive evaluation in hundreds of monkeys at Baylor College of Medicine, and the Division of Biological Standards at the NIH, the Sabin strains were chosen for licensure primarily on the basis of lower neurotropism, but also based on genetic stability on passage in humans and a lower ability to spread to contacts (reviewed in Sutter et al. [4] and Furesz [22]). Efforts to eradicate polio and to prevent reemergence are presented in detail in the following section. Initial paradigms attributed to the different properties of the individual vaccines have not always held true in all circumstances [23].

## The Epidemiology of Polio

Epidemiological studies to discover the means of preventing a disease usually begin with the recognition of a new pattern of similar symptoms among those affected and the establishment of a case definition. Discovering means for preventing the disease may start before the disease agent is discovered and characterized, but is certainly accelerated once this characterization becomes available together with the means of quantifying intervention strategies. It is much less



common to start with an agent and then search for a disease as in the case of human anelloviruses [24]. Human anelloviruses are small circular DNA viruses considered to be orphan viruses. They were initially discovered in a patient with hepatitis, but subsequent research indicated no causal link to hepatitis and it has been very difficult to associate them with any other specific disease. However, this section of the review will start with a description of those physical aspects of poliovirus that have the most impact on epidemiology of the disease. This is because there is already a clear case definition for polio and poliomyelitis, polioviruses have been recognized as the causative agents of these diseases, numerous methods for characterizing poliovirus and preventing poliovirus infections have been developed and tested, and the disease is approaching elimination or eradication.

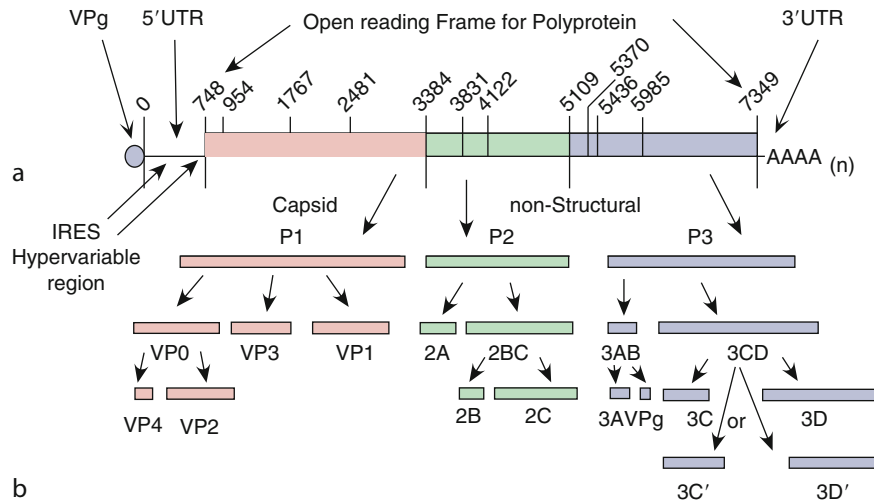
### Structural and Functional Organization of the Poliovirus Genome

Polioviruses belong to the *Picornaviridae* virus family. The *Picornaviridae* genome consist of a single strand of positive-sense RNA approximately 7,500 nucleotides located within a protein capsid made up of 60 capsomeres that forms a virion 27–30 nm in diameter. The genome is organized from its 5' end to its 3' end into a number of functional regions (Fig. 3) that include a 5' untranslated region (5'UTR) that regulate translation and replication [25], a long open reading frame that encodes a single large polypeptide that is cleaved after translation into four structural capsid proteins and a number of nonstructural proteins including an RNA polymerase, and a short 3' untranslated region that is attached to a poly-A tail in both viral mRNA and genomic RNA in the virion [25] (see reviews by Wimmer et al. [26], Racaniello [25], and Sutter et al. [4]). The positive-sense single strand of genomic RNA in the virion, serves directly as an mRNA template for translation to viral proteins once the virion penetrates its host cell membrane. Later it serves as a template for synthesis of a complimentary negative sense strand. The current understanding of the physical and genetic aspects of polio was greatly facilitated by the development of and the current commercial availability of methods for easily extracting viral nucleic acids from poliovirus and poliovirus-infected cells

and analyzing and manipulating these sequences. Some of these studies led to the unanticipated conclusion that poliovirus capsid proteins and the sequences that encode them define polioviruses, whereas all other elements in the poliovirus genome may be substituted by genomic recombination with equivalent sequences from closely related isolates of enterovirus species C in vivo and even more distantly related rhinoviruses in the laboratory as long as functionality is maintained (reviewed by Kew et al. [3]). Finally, advances in molecular biology also enabled poliovirus to be the first virus to be synthesized from nucleotides in a test tube [27, 28].

The 5'UTR was first subdivided into a highly conserved region (nucleotides 1–650) and a hypervariable region (nucleotides 651–750) based on an analysis of 33 wild-type 3 polioviruses [29]. A series of stem-loop structures with a high degree of secondary structure were proposed to be present within the conserved region by Pilipenko et al. [30] and Skinner et al. [31]. A single nucleotide substitution in a loop structure in stem-loop V of the 5'UTR significantly influenced the neurovirulence of poliovirus isolates from all three serotypes and affected the maximum temperature at which viral isolate replicate efficiently (see reviews by Kew et al. [3] and Sutter et al. [4] and discussions on poliovirus evolution starting on page 8137). The hypervariable region appears to be much less structured, reflecting the high degree of variation and the U nucleotide richness [29].

An Internal Ribosome Entry Site [32, 33], IRES, enables uncapped RNA from *Picornaviridae* to be translated in eukaryotic cells by host ribosomes [25]. One of the first steps in initiation of viral translation is the binding [34] of cellular RNA binding proteins PCB<sub>1</sub> and PCB<sub>2</sub> to stem-loop IV of the IRES. This enables the 40S ribosomal unit to bind to the IRES and continue the process of translation as if the RNA was a capped eukaryotic mRNA. Functional IRES elements can be interchanged among *Picornaviridae* [3]. Nucleotide differences in the conserved 5'UTR among different isolates were unevenly distributed [29] with changes tending to conserve the stem structures. In contrast, the hypervariable region did not seem to have a highly conserved secondary structure and nucleotide differences appeared to be more or less evenly spread throughout [29]. While the length of the hypervariable



b

Serotype	Nucleotide	Gene	Amino Acid	Temperature Sensitivity
Sabin 1	A480G	5'UTR	Non-coding	Yes
	G935U	VP4	ala 65 ser	-
	U2438A	VP3	lys 225 met	-
	G2791A	VP1	ala 106 thr	-
	C2879U	VP1	leu 134 phe	-
	U6203C	3D	tyr 73 his	Yes
Sabin 2	G481A	5'UTR	Non-coding	Yes
	C2909U	VP1	thr 143 iso	-
Sabin 3	C472U	5'UTR	Non-coding	Yes
	C2034U	VP3	ser 91 phe	-
	U2493C	VP1	iso 6 thre	Yes

c

### Polio and Its Epidemiology. Figure 3

Organization of the polio viral genome, posttranslational processing of the nascent poliovirus polyprotein, and the nucleotide substitutions that differentiate attenuated oral polio vaccine strains from their neurovirulent progenitors. The RNA positive-sense strand genome of Sabin 2 based on GenBank/EMBL/DDBJ entry AY184220 (a) is covalently linked to the viral encoded protein VPg. There is a single open reading frame flanked by a 5' and a 3' untranslated sequence (UTR). An internal ribosomal entry site (IRES) in the 5'UTR allows the uncapped polio genomic RNA to serve as mRNA for translation on host cell ribosomes. The open reading frame is translated into a single poliovirus polyprotein that undergoes a series of posttranslational proteolytic cleavages (b) while it is still being translated. Some of the intermediate products have enzymatic and/or structural functions that differ from those of the final cleavage products. Poliovirus genomic and mRNA terminates in a poly-A tail. The attenuation of neurovirulence in Sabin 2 and the other 2 serotypes, Sabin 1 (GenBank/EMBL/DDBJ entry V01150) and Sabin 3 (GenBank/EMBL/DDBJ entry X00925), of poliovirus strains used for the live polio vaccine result from the nucleotide and amino acid substitutions shown in (c). Reversion of these substitutions may restore a neurovirulent phenotype for the progeny of these vaccine strains. Nucleotide substitutions are indicated by the original nucleotide of the parental strain the nucleotide position, and the substituted nucleotide in the vaccine strain (Adenine Uracil, Guanine, or Cytosine). Amino acid substitutions are indicated by the parental amino acid, the position of the amino acid in the final cleavage product, and the amino acid in the vaccine strain (*alanine, histidine, isoleucine, leucine, methionine, phenylalanine, serine, threonine, and tyrosine*). ((b) Based on: [1] Krausslich HG, et al. [37] and [2] Kitamura N, et al. [290]. (c) Modified from: Kew OM, et al. [3])

region was generally conserved suggesting an unknown function [29], small deletions were tolerated [35].

*Picornaviridae* have a genome of approximately 7,200–7,400 nt with a single open reading frame (ORF). While this ORF encodes four capsid proteins and at least seven viral proteins (Fig. 3), these proteins are only produced after the initial translation product, a single polypeptide, is enzymatically cleaved into smaller and smaller polyproteins during and after translation (posttranslational processing). The polypeptide is cleaved in an ordered series of steps (Fig. 3b), by viral encoded protease activity within the nascent polypeptide (self-cleavage) and *in trans* from viral proteases released after cleavage. Interestingly some of the intermediate cleavage products have unique activities by themselves that contribute to the replication cycle of the virus, but which differ from those of the final cleavage products (reviewed by Racaniello [36] and Krausslich et al. [37]). Properties of the polioviral capsid proteins define the epidemiology of polioviruses. The most important aspects of the structure of the four capsid proteins, their assembly into capsomeres and organization within adjacent capsomeres that relate to the epidemiology of polio, will be discussed above on page 8131. The 900–906 nucleotide sequence of the VP1 of polioviruses has become the minimum standard for determining the evolutionary relationship among polioviruses and the rate at which they evolve [4, 38, 39].

Many of the nonstructural proteins and intermediate cleavage products are multifunctional and act at a number of steps in RNA synthesis (reviewed in [4, 25]). Most of the nonstructural proteins will only be mentioned in passing since equivalent nonstructural proteins from other related picornaviruses may replace all of the nonstructural viral proteins as long as functional sites including cleavage recognition sites are maintained (reviewed in [3]). The resultant chimeric recombinants behave as polioviruses. One nonstructural protein, the RNA polymerase, will be discussed in some detail (see page 8135) because of its profound effect on polio epidemiology regardless of its source.

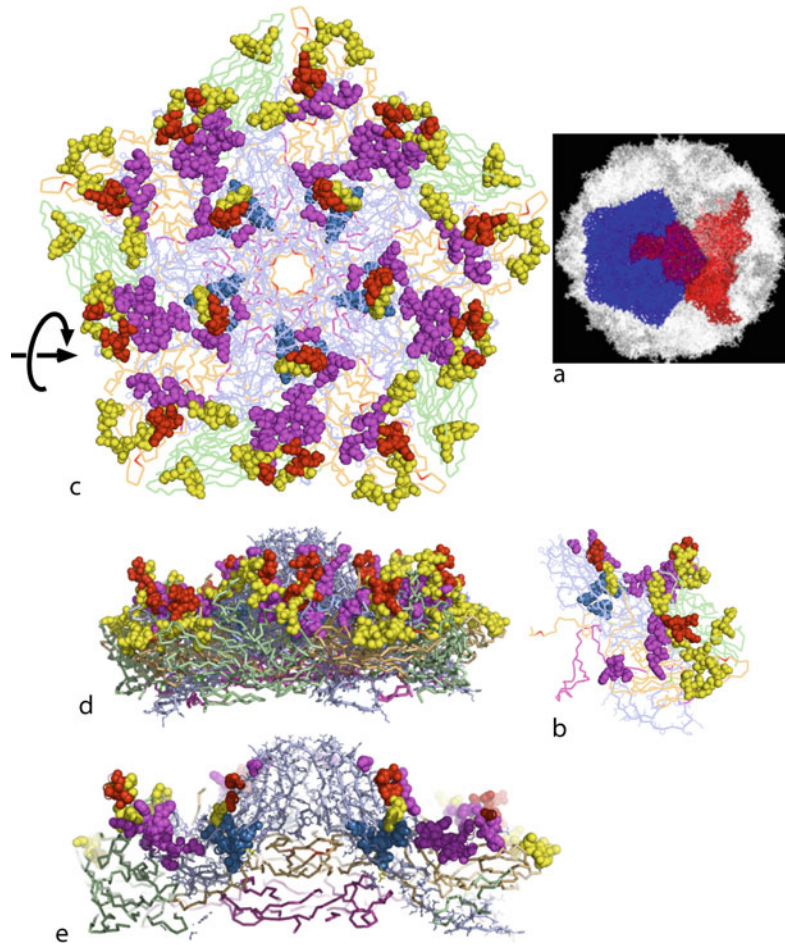
The secondary structure in the 3'UTR that may play a role in translation and replication of picornaviral genomic RNA has been reviewed [25]. A nucleotide difference between Sabin serotype 1 and its wild parent

influences the temperature at which serotype 1 can replicate [40]. A poly-A tail is present on both genomic and mRNA that stimulates the cap-independent, internal ribosome entry site (IRES)-driven translation of poliovirus RNA in a mammalian cell-free system by tenfold [41].

Both genomic and minus strand RNA are linked to the small viral encoded protein, VPg, (Fig. 3a) through pUpU bound to tyrosine, the third amino acid from NH terminal end of VPg, by a phosphodiester bond [42]. VPg is also present in infected cells in an unmodified form and bound to pUpU through the same 0<sup>4</sup>-phosphotyrosine bond found in the covalently linked forms [42, 43]. The uridylylation of VPg takes place on the opposite side of the polymerase that binds RNA. A host encoded unlinking enzyme that cleaves the 0<sup>4</sup>-phosphotyrosine bond between VPg and RNA has been described [44] although its role in replication has not been established. The poliovirus encoded VPg can be replaced by VPg from echoviruses [43].

A mature infectious poliovirus consists of a single sense strand of polyadenylated RNA covalently linked to a viral encoded protein, VPg, surrounded by an icosahedral protein coat, the capsid, made up of 60 capsomeres that each contain a single copy of each of the four viral capsid proteins. Adjacent capsomeres are organized around both fivefold and threefold axes of symmetry and the surface around these axes is organized into a series of regular protrusions and depressions (Fig. 4). The capsid structure is metastable [45] rather than rigid and internal parts of capsid may even be transiently expressed on the surface ([46], review in [25]) exposing additional epitopes such as PALTAVE inVP1 [47].

Capsid proteins are the first viral encoded proteins to appear on the nascent poliovirus polyprotein and are cleaved from the nascent polyprotein into an intermediate polyprotein, P1, by 2A<sup>P<sup>ro</sup></sup> while the full-length polyprotein is still being synthesized. P1 is processed into final cleavage products VP1 and VP3 and an intermediate cleavage product VP0. VP0 is only cleaved into VP2 and VP4 during the final stages of maturation of the virion. The protein chains of VP1, VP2, and VP3 are arranged in wedge-like structures with extruding loops that interact to form the major (NAGIa, NAGIIa, and NAGIIIa) and minor (NAGIb, NAGIIb, and NAGIIIb) neutralizing antigenic epitopes [48]. Amino



**Polio and Its Epidemiology. Figure 4**

The hydrophobic pocket and amino acid residues in the neutralizing antigenic epitopes and receptor binding sites of the Sabin 2 polio vaccine strain. The three-dimensional structures represent capsomeres 1–5 from human serotype 2 poliovirus, Genbank/EMBL/DBJ entry 1eah. The backbones of the amino acid chains of the capsid proteins are represented by *light blue, pale green, light orange, and magenta colored ribbons* for VP1, VP2, VP3, and VP4, respectively. Amino acid residues at the surface of the hydrophobic pocket are represented by *blue spheres*. Amino acid residues within the epitopes recognized by neutralizing antibodies are represented by *yellow spheres*, those involved in receptor recognition and binding are represented by *magenta spheres*, and amino acid residues shared by both antigenic sites and receptor binding sites are represented by *red spheres*. The figure was prepared using the MacPymol program (DeLano Scientific LLC, [www.pymol.org](http://www.pymol.org)). Figure (a) is a representation of the entire capsid of poliovirus showing the positions of the threefold (in *red*) and fivefold (in *blue*) symmetrical organization of the capsomeres. Each poliovirus capsomere (b) contains a single copy of each of the four viral capsid proteins. Five capsomeres are assembled around a fivefold axis of symmetry shown in (c) and by *blue* in (a). They also assemble around a threefold axis of symmetry shown in *red* in (a). Figure (c) represents an external view of the five capsomeres at the fivefold axis of symmetry. Figure (d) is the side view of the same five capsomeres formed by rotating the figure in (c) in the direction of the *circular arrow*, so that the lower structures in (c) are nearest the viewer and the internal surfaces of the capsid proteins are facing downward. Figure (e) is a transverse section of the figure in (d) at the position of the *straight arrow* in (c) to more clearly illustrate the topography of the surface of the virion. An animated “Interactive 3D Complement” (I3DC) for the structures in this figure appears in *Proteopedia* at [http://proteopedia.org/w/Polio\\_Epidemiology](http://proteopedia.org/w/Polio_Epidemiology)

acid differences within neutralizing antigenic sites divide polioviruses into three serotypes with limited cross-reactivity [49]. The amino acids of the neutralizing antigenic sites have been mapped onto the three-dimensional structures of the viral capsid of type 2 poliovirus as colored spheres, Fig. 4. Those that are unique to the neutralizing antigenic epitopes are colored yellow. Some amino acid residues in and adjacent to these neutralizing antigenic sites (red spheres in Fig. 4) are also involved in receptor binding and this may have restricted the number of serotypes [50] and influenced evolution in these epitopes in the absence of immunoselection especially during the emergence of vaccine-derived polioviruses (VDPVs) (see page 8159).

The three-dimensional view of the structure of the capsomeres at the fivefold axis of symmetry reveals an elevated central plateau with a hole in the middle surrounded by a depression called the canyon [51, 52]. The fivefold axis of symmetry for type 2 poliovirus is shown in Fig. 4. A number of conserved amino acids and amino acids within and adjacent to the serotype-specific neutralizing epitopes are located on the surface of the canyon walls and have been implicated in interaction with the poliovirus receptor [26, 50].

The human encoded, poliovirus receptor, CD155, belongs to the immunoglobulin super gene family and has one variable and two constant immunoglobulin-like domains (residues 28–337) [11]. This human encoded gene has alternative splice sites that result in two membrane-bound and two secreted isoforms [53]. The variable domain 1 penetrates the canyon and binds to amino acid residues from all three external capsid proteins and the principal binding sites are at the bottom of the canyon above the hydrophobic pocket (blue spheres in Fig. 4) and on the outer side of the canyon rim [50, 54]. The residues of type 1 poliovirus involving receptor virion binding include residues 102–108, 166–169, 213–214, 222–236, 293–297, 301–302 in VP1, residues 140–144, 170–172 in VP2, and 58–62, 93, and 182–186 in VP3. The equivalent residues for serotype 2 poliovirus have been mapped onto the three-dimensional capsid structure as red (shared with neutralizing antigenic epitopes) and magenta spheres for those associated only with the receptor binding sites (Fig. 4). Cryo-electron microscope studies have shown the binding of the poliovirus

to the virion to be a two-step process [54]. The initial binding of the receptor to amino acid residues along the canyon wall results in little or no change in virion structure. However, this binding rapidly sets into motion conformational changes leading to the 135 S or A particle state that initiates uncoating and the start of the infections cycle [45, 54].

The human poliovirus receptor has been cloned and used to establish a murine cell line, L20B, where expression of the poliovirus receptor allows infection and growth of polio from clinical and other samples but not most other human non-polio enteroviruses [13, 55, 56]. Transgenic mice, PVR Tg-21 mice that express the human poliovirus receptor, not only support poliovirus infection and present with neurological symptoms, but allow determination of the relative neurovirulence of the isolates [12, 57–59].

A hydrophobic pocket (blue spheres in Fig. 4) located below the canyon floor is normally occupied by pocket factors such as sphingosine-like molecules including palmitic and myristic acids and hydrophobic compounds, that stabilize the capsid, enable receptor docking and whose removal is a necessary prerequisite for uncoating [25, 45, 54, 60].

Small molecules such as pleconaryl and isoflavones can bind in this hydrophobic pocket and exert antiviral effects by affecting the binding of the receptor or enhancing the stability of the virion and preventing uncoating [25]. Because of the metastable nature of the capsid, mutations distant from the receptor and drug binding sites can compensate mutations in the respective binding sites [45, 61].

Molecular analysis studies of isolates shed during persistent infections of immunodeficient patients [62, 63] and from phylogenetically related aVDPVs from environmental samples help pinpoint amino acid substitutions in capsid proteins that determine antigenicity, receptor recognition, attenuation of neurovirulence, and properties of the hydrophobic pocket. Other changes, some of which are at interfaces between the threefold or fivefold interfaces of capsomeres may also affect these properties indirectly.

In order for single stranded, positive-sense genomic RNA to be incorporated into progeny of the infecting virus, a complimentary negative RNA strand must first be synthesized using the original single stranded positive-sense RNA genome as template, and this

complimentary negative strand must then be used as template for synthesis of new positive-strand RNAs. While some positive-sense copies are incorporated into progeny virions as genomic RNA, other newly synthesized positive-strand RNAs serve as templates to repeat and amplify RNA replication and/or for translation to produce more viral proteins. Eukaryotic cells that serve as the host for poliovirus replication lack a polymerase that can synthesize complimentary RNA from an RNA template. Therefore the virus must encode its own polymerase. Since the single stranded positive-sense RNA genome of the infecting virion is also an mRNA that is immediately translated, the virion does not have to incorporate the polymerase into the virion itself to start replication. The translation product of the 3D<sup>pol</sup> gene (Fig. 3) is the required RNA-primed RNA polymerase. Both the intermediate cleavage products that contain the 3D<sup>pro</sup> and the final cleavage product are multifunctional and the crude replication complex also contains other viral proteins and protein cleavage intermediates such as 2BC, 2C, and 3AB as well as host proteins (reviewed in [4, 25]). The binding site for RNA template and primer are on one face of 3D<sup>pol</sup> and a binding site for the uridylylation of VPg, a prerequisite for covalent linking of VPg to viral RNA, is on the opposite face [25].

One important contrast between genomic DNA replication in eukaryotic host cells and genomic RNA replication in *Picornaviridae* relates to the fidelity of replication. Specifically, there is an elaborate proof-reading mechanism combined with pathways for correcting misincorporations during replication of Eukaryotic DNA that is lacking in the RNA-primed RNA polymerase complex for viral replication [64, 65]. This leads to such a high evolutionary rate for polioviruses that it borders on error catastrophe [66, 67] (discussed further below).

### Poliovirus Infections in Cells, Individuals, and Populations

This section will deal with the epidemiological aspects of poliovirus infections at three levels, infections in single cells, infections in a single individual, and infections in populations of individuals. The normal infectious cycle of a poliovirus starts with recognition and attachment to poliovirus-specific cell receptors on

susceptible cells of human or closely related primate origin. It continues with penetration and uncoating, translation of viral RNA, posttranslational processing of viral polyproteins, replication of viral genomes, assembly and maturation of progeny viruses culminating in the release of infectious polioviruses. During this process, the virus employs and modifies host cell functions to optimize viral yield. The observation that viral RNA and cDNA is infectious when transfected into permissive host cells has allowed recovery of virus from extracted genomic RNA, cloned cDNA or RNA translated from cloned DNA [68–70], genomic RNA immobilized on FTA paper (WHO 16th Informal Consultation Of The Global Polio Laboratory Network, September 2010, Geneva, Switzerland), and from polioviral RNA synthesized in a test tube from individual nucleotides [27, 71]. The infectious cycle has been reviewed extensively. The reader is referred to the following reviews for further reading [4, 25, 72].

### Poliovirus Infections at the Level of the Host Cell

All polioviruses recognize a single host cell receptor [50], CD155, also known as the poliovirus receptor (PVR). Identification and cloning of the poliovirus receptor CD155 [11] allowed the creation of cell lines [13] and animal models [58, 59] for the study of polioviral infections in non-primate hosts. The interaction of virion and receptor is complex [25]. The capsid structure is dynamic allowing the transient presence of internal portions and epitopes of capsid proteins on the outer surface of the virion [25, 46] including the N-terminus of VP1 even before uncoating. The shape of the receptor and its position relative to the host cell membrane and the canyon on the virion into which it fits bring the fivefold axis of capsomeres in close proximity to the cell membrane [54].

Conformational changes, induced shortly after the virion–receptor interaction, are required to initiate the uncoating process (reviewed in Racaniello [25]). The capsid begins to disassociate during a transition to the A particle. The A particle contains the viral RNA but has lost its VP4 capsid proteins. The N-terminal of the VP1 externalizes and may insert into the plasma membrane. Viral RNA is believed to enter the cell at or near the fivefold axis through a continuous channel formed in part by VP1 that continues through the cell

membrane [54]. VP4 plays a part in formation of the pore. The pore for poliovirus entry is probably not formed within endosomes [25]. Small molecules that sit in the hydrophobic pocket may influence these conformational changes without affecting receptor binding [61, 73–75].

The only viral proteins in the virion are the four capsid proteins and the VPg covalently linked to the genomic RNA. The internal ribosomal entry site (IRES) on uncoated polioviral RNA enables translation of the viral polyprotein on host cell ribosomes (reviewed in [3, 25]). VPg appears to be cleaved from this RNA and subsequently synthesized viral RNA that will be used as mRNA [44]. Nuclear trafficking of cellular proteins is downregulated shortly after infection resulting in accumulation of host nuclear proteins in the cytoplasm that could function alone or in combination with viral encoded proteins in viral RNA translation, synthesis, and packaging [73]. Downregulation may be due in part to specific degradation of two host transporters, Nup 153 and p62. A full-length polyprotein is not observed in spite of being encoded by the single long open reading frame since posttranslational cleavage of the polyprotein is initiated as soon as the portion encoding the 2A<sup>Pro</sup> has been translated. Many of the nonstructural proteins and intermediate cleavage products are multifunctional and act directly or indirectly at a number of steps in the RNA synthesis pathway (reviewed in [4, 25]). One example is the aforementioned viral encoded protease, 2A<sup>Pro</sup>, that also shuts off host protein synthesis by cleaving eIF4G. eIF4G is required for translation of capped eukaryotic mRNAs, while the C-terminal of the cleavage product enhances IRES activity [25]. 2A<sup>Pro</sup> is also important for negative strand but not positive-strand RNA synthesis [76]. Another example is the intermediate cleavage product, 3CD<sup>Pro</sup>, that also participates in the posttranslational processing of the polyprotein.

The last protein of the polyprotein to be translated is the 3D polymerase. RNA-primed RNA synthesis is initiated once the 3D has been released from the polyprotein reviewed in [25]. VPg-pUpU or VPg itself could act as a precursor for RNA synthesis by hybridizing to template RNA [44, 77]. The binding site for template and primer are on one side and that for VPg is on the other side. A replicate intermediate is formed

and consists of a positive-sense RNA with 6–8 nascent negative strand RNAs. The negative sense strand serves in turn as template for synthesis of a 30-fold excess of new sense strand RNAs. Full-length dsRNAs can be isolated from infected cells. Altogether the genomic RNA is amplified up to 50,000-fold. VPg is bound to both genomic RNA and negative sense RNA.

Poliovirus and other picornaviruses employ a quasispecies reproductive strategy [64, 78] where the lack of proofreading rapidly results in a mixture of progeny with modified genomes containing randomly positioned single nucleotide substitutions. Genomic recombination is a second method of evolution where a single event results in substitutions of many nucleotides from a different poliovirus or closely related non-polio enterovirus for the equivalent sequence in the original poliovirus. The majority of single nucleotide substitutions are deleterious or neutral; however some may confer a reproductive advantage for progeny for growth in the current or future host and/or for host-to-host transmission. Evolutionary changes become “fixed” by selective outgrowth of individual members of the quasispecies that pass through bottlenecks within and between hosts [79]. Two evolutionary pathways, the very high number of progeny (>10,000 per infected cell) and outgrowth by chance selection and/or a selective advantage, result in one of the highest observed rates of molecular evolution [39].

RNA is synthesized from four nucleotides, two pyrimidines (uracil and cytosine) and two purines (adenine and guanine). The most common route for polioviral evolution is by nucleotide misincorporation (single nucleotide substitution) in the absence of both proofreading and post-incorporation excision–repair pathways. The nucleotide position that is substituted is probably random but may be influenced to some extent by secondary structure and the adjacent nucleotides. Quasispeciation arises from the fact that the remaining progeny retain the original nucleotide at the position of each unique substitution in an individual progeny virus, while within the cloud of progeny each isolate may have a unique substitution at a different position in the genome.

Among the isolates that make up the quasispecies, substitutions should be found at each position in the genome at an equal frequency, at least in theory.

However, substitutions are much more frequently observed in some positions than in others. Two related factors contribute significantly to the nonuniform distribution (see page 8140) of observed substitutions along the genome. The first is that almost all observations have been made with RNA extracted from the quasispecies that arose during replication of a viable virus directly in the primary host or after amplification of one or more isolates from the quasispecies *in vivo* in a second host or *ex vivo* in tissue culture. The second is that substitutions in some positions produce nonviable or less fit offspring that are eliminated during this amplification process.

Sequence-specific variability, based on the individual nucleotide base and its nearest neighbors [67], and inherent characteristics of the polymerase are other factors that contribute to the nonuniform distribution. If misincorporations were unbiased, transversions (the substitution of a pyrimidine by a purine or vice versa) would be expected to occur at twice the rate of transition (the substitution of a purine with a purine or a pyrimidine with a pyrimidine). However empiric observations have revealed a polymerase-based bias of approximately ten to one in favor of transitions [39]. To currently include sequence data from the genomes of nonviable progeny requires either amplification of individual genomes by a process that does not require an active poliovirus infection but that includes high fidelity with proofreading and excision-repair (reverse transcribing the genomic RNA and cloning the cDNA of all viable and nonviable poliovirus progeny into plasmids that can be amplified in bacterial strains with high fidelity, proofreading, and error correction) or by direct sequencing of individual genomes without amplification (chip/array sequencing technology) [66, 80]. Neither approach is currently very easy to apply since both would require individually processing large numbers of genomes equivalents, although Crotty et al. were able to calculate a rate of  $2.1 \times 10^{-2}$  substitutions per site by direct measurement of mutations in the VP1 of 55 cloned genomes after a single cycle of *in vitro* virus growth. Massive parallel next-generation sequencing may offer the best approach for analyzing viable and nonviable members of a quasispecies [289].

Wild poliovirus genomes frequently recombine (recombination) with polioviruses and closely related

non-polio enterovirus genomes [81, 82]. This recombination can only occur during concurrent infection of a single cell by both parental isolates. Intratypic recombination may occur even within capsid proteins [83–85].

The noncapsid regions of polioviruses are most similar in sequence to other members of the enterovirus C genotype that includes Coxsackie A virus (CAV) serotypes 1, 11, 13, 15, 17, 18, 19, 20, 21, 22, and 24, and these sequences are readily shuffled among polioviruses and the other members of this group [86, 87]. In fact polioviruses show evidence of having evolved from C-cluster Coxsackie A viruses and may reemerge from them after eradication [87]. Interspecific recombination contributes to the phenotypic biodiversity of polioviruses and may favor the emergence of circulating vaccine-derived polioviruses, cVDPVs [88].

Recombination is not site specific, does not require extensive homology between genomes at the crossover site, and most likely occurs by an exchange of templates by the synthesis of complimentary RNA by the RNA-primed RNA polymerase rather than by breakage rejoining [89]. Intratypic (same serotype) and intertypic (different serotype) recombination *in vitro* occurred at  $1.3 \times 10^{-3}$  and  $7.6 \times 10^{-6}$ , respectively [89], while recombinations between polio and NPEVs occurred at a frequency of  $10^{-6}$  [87]. Administration of trivalent oral vaccine anywhere and in areas where wild polioviruses and genotype C viruses co-circulate provides the conditions for concurrent infections and polio vaccine-polio vaccine, polio vaccine-wild polio, and polio vaccine-NPEV, as well as endemic wild polio-NPEV recombinations. For examples of such recombinations see molecular analyses of isolates from the cVDPV outbreaks in Haiti and Dominican Republic [90] and Indonesia [91] and in individual cases [92].

Molecular epidemiology is the study of disease and factors controlling the presence or absence of a disease or pathogen using molecular data (DNA, RNA, or protein sequences). The next portion of this section will concentrate on those aspects of molecular epidemiology that impact on the epidemiology of polio.

Polioviruses and other enteroviruses are among the organisms with the highest rate of misincorporation ([67, 78], and reviewed [39]). Misincorporation comes at a high cost, namely, only approximately 10% of



the >10,000 progeny from a single infected cell are viable [67]. This high frequency of misincorporation helps to explain the high ratio of physical to infectious particles [93]. Studies with ribavirin [66, 94], an antiviral drug acting as a nucleoside analogue, have shown that the misincorporation rate of polioviruses is close to the catastrophe error rate, that is, the transition point where a modest increase in misincorporation results in a drastic decrease in viability. In these experiments a 9.7-fold increase in mutagenesis resulted in a 99.3% loss in viral genome infectivity after a single round of replication, while a less than twofold increase in the natural mutational frequency resulted in a 50% loss of viability.

Fitness is based on the overall performance during viral replication [67], a complex process, involving recognition of and binding to the host cell, uncoating/entry, initiation of protein synthesis before RNA replication takes place, regulation of replication and translation once RNA replication is initiated, culminating with assembly, maturation and externalization of mature virus and survivability until subsequent infection of the next host cell or organism. Changes can affect more than one of these processes. For example, an increase in the mutational rate in infections in the presence of the nucleoside analog ribavirin not only led to an increase in nonviable genomes, but also caused a reduction in the total number of viral genomes produced [66]. One of the advantages of the quasispecies nature of poliovirus offspring is that isolates with a selective advantage to new growth conditions may already exist in the population [66]. Studies on mixed infections in PVR Tg21-transgenic mice suggest that random selection may play a role in the selection of which genomic variants within a mixed infection in the gut infect the CNS since virus isolated from the CNS was not always the most neurovirulent [95]. Other experiments showed that increased fidelity of the polymerase reduced viral fitness in the PVR Tg21-transgenic mice [96] or in tissue culture [97]. An alternate explanation for selection was proposed by Andino and colleagues [98]. They provide evidence that the quasispecies is not just a collection of individual variants, but a group of interactive variants and that fitness and selection may occur at the level of the population rather than at the level of individual genomes. In their study, an increase in polymerase

fidelity affected viral adaptation and pathogenesis in addition to genome variability. Data supporting the suggestion that minor components can alter the phenotype of quasispecies comes from retroviral infections [99] and studies with VSV [100].

A number of studies have shown that single nucleotide misincorporations by the polio RNA-primed RNA polymerase accumulate at a more or less constant rate that can be used as a “molecular clock” to estimate evolutionary time between isolates and to determine whether sequence differences between two polioviruses isolated within a given time interval are consistent with a shared, direct evolutionary pathway between them [38, 39, 101, 102]. In general, the rate of accumulation and fixation of single nucleotide substitutions appears to be similar for all isolates regardless of kind (all three serotypes of wild, vaccine, or vaccine-derived polioviruses), type of polymerase (original intact polymerase or chimeric or complete recombinant from the same serotype, a different serotype or even a group C non-polio enterovirus), or type of infection (transient in immune competent individuals, persistent in immunodeficient patients, or even in the very elderly where waning immunity may play a role in selection) [4, 39, 79, 90, 101, 103–110]. Moreover, the rate of third codon position synonymous substitutions appeared to be fairly constant throughout the period of virus excretion in a persistently infected individual [107]. Using molecular observations from full-length genome sequences from viruses isolated during a 10 year long outbreak established from a single imported founder virus, Jorba et al. [39] calibrated five clocks based on five different classes of nucleotide substitutions. The constants for total substitutions ( $K_t$ ), synonymous third position substitutions in coding regions ( $K_s$ ), synonymous transitions ( $A_s$ ), synonymous transversions ( $B_s$ ), and non-synonymous substitutions ( $K_a$ ) were  $1.03 \pm 0.10 \times 10^{-2}$ ,  $1.00 \pm 0.08 \times 10^{-2}$ ,  $0.96 \pm 0.09 \times 10^{-2}$ ,  $0.10 \pm 0.03 \times 10^{-2}$ , and  $0.03 \pm 0.01 \times 10^{-2}$  substitutions/site/year, respectively. The rates were similar whether calculated using linear regression, a maximum likelihood/single-rate dated tip method, and Bayesian inference. The first two constants were mostly controlled by the third. As for saturation, third position synonymous transitions become evident by 10 years and complete saturate within 65 years while saturation of synonymous transversions was predicted

to be minimal at 20 years and incomplete even at 100 years. This wide variation in calculated time constants depending on the type of substitutions together with differences in the estimated time until all possible changes become saturated, provides a flexibility that allows one or more clock to be applied to characterize the range from evolution in outbreaks between very closely related isolates with short intervals between isolations, to comparison between much more distantly related polioviruses or related enteroviruses. It is interesting to note that the molecular clocks are fairly constant given that intratypic and intrageneous recombination can result in complete or partial substitution of the polymerase whose intrinsic properties presumably govern the rate of misincorporation. Mutations may increase non-synonymous mutation rates [111] while others decrease them [96]. Multiple recombination events [83, 85] must be ruled out or taken into account when calculating time clocks based on the number of nucleotide differences.

Different factors that affect fitness and determine the viability of individual viral offspring result in differences in observed substitution rates and patterns in the different functional elements of the genome shown in Fig. 3. Namely, using the rate of substitutions in the VP1 capsid protein as reference, the rates of substitutions are approximately half in the conserved region of the 5'UTR, approximately threefold higher in the hypervariable region of the 5'UTR, and equivalent or somewhat lower in the remainder of the ORF [4, 39, 101, 109]. The data for nucleotide substitutions in the nonstructural P2 and especially the P3 regions of the ORF and the 3'UTR are less accurate and less informative due to frequent recombinations among polioviruses and between polioviruses and non-polio enteroviruses within these regions.

The genetic code introduces a bias in the position of observed substitutions. Substitutions in the third position of a codon are least likely to result in an amino acid change and these synonymous substitutions are by far the most abundantly observed in wild poliovirus, polio vaccine and VDPV infections [4, 79]. Non-synonymous substitutions that occur in the initial stages of vaccine infections restore replicative fitness and in many cases neurovirulence [3, 112] while those in persistent infections may influence receptor-virus interaction (see page 8133, 8159).

Three-dimensional requirements also bias the observed distribution of substitutions. Maintenance of stem of the stem-loop structure in the conserved region of the 5'UTR especially within the IRES appears to be one of the major constraints on viability. For example, complimentary paired double substitutions that maintained stem structure were frequently found in the loop V of evolutionarily related environmental isolates [109] whereas a single nucleotide substitution in a loop of Loop V is a dominant determinant of attenuation of neurovirulence and growth at elevated temperatures. Other examples of three-dimensional effects are mutations that occur at distances from functional sites that influence the viral response to antiviral drugs in the hydrophobic pocket [61] and mutations that occur at the interfaces between proteins and at the N-terminals of VP1 and VP4 that may affect structural stability and the receptor-induced transitions [45].

Due to the complex nature of polioviral replication and multi-functionality of viral enzymes and viral three-dimensional structures, selective pressures that operate on one structure or function may affect another seemingly unrelated property. One of these apparent paradoxes is the fact that some RNA viruses including poliovirus may diverge antigenically in the absence of immune selection [113]. One of the features that distinguishes polio vaccine evolution during persistent infections in total B-cell deficient immunodeficient patients from evolution during person-to-person transmission in immune competent but naïve individuals is that isolates from the former but not the latter have high numbers of amino acid substitutions in and around neutralizing antigenic sites [3, 4, 62, 114]. Antibody titers tend to wane in the elderly. The finding of amino acid substitutions at or near neutralizing antigenic sites during infection of the elderly with type 1 monovalent mOPV [103] may suggest that waning immunity may create a situation resembling the early stages in establishment of persistence in immunodeficient patients. Since some of the amino acids and structural organizations are shared by neutralizing antigenic sites and receptor binding sites, the high mutation rate in neutralizing antigenic sites is more likely the result from selective pressures governing receptor-virus interaction during establishment and maintenance of persistence than on non-humoral mediated immune selection or selection by variations

in anti-polio antibodies in the IVIg regimens these immunodeficient patients receive to compensate for their B-cell deficiencies. This sharing of functions has also been suggested to be one of the reasons why there are only three serotypes of poliovirus [26, 50]. It is commonly accepted that the evolution of a fourth serotype would require receptor switching of a non-polio enterovirus to the use of the PVR, CD155. A somewhat paradoxical alternative for emergence of a fourth serotype may be through antigenic evolution during persistent infections in immunodeficient individuals as a result of selective pressures relating to receptor binding. Consistent with this is the observation that cohorts of immunized individuals who had high titers against vaccine strains had significantly reduced geometric mean titers against highly diverged neurovirulent vaccine-derived viruses that were isolated from environmental samples [109, 115] and individual titers against some of these isolates were <1:8 in 7% of the adults [109, 116].

Any discussion on molecular evolution of polioviruses and their effects on polio epidemiology would be incomplete without a discussion on vaccine-derived polioviruses, VDPVs (see page 8157 and the section on future directions). Evolution of live attenuated polio vaccine occurs by the same processes as in wild polioviruses, namely, by accumulation of single nucleotide substitutions and through genomic recombination. Evolution of VDPVs occurs during person-to-person circulation in cohorts of naïve or under-immunized individuals especially after interruption of vaccination, or during persistent infections of immunodeficient individuals [3, 80, 106, 114, 117, 118]. The letters “c” or “i” for viruses that evolved during person-to-person circulation or during persistent infections of immunodeficient individuals, respectively, are appended before “VDPV” when the evolutionary pathway is known. The prefix “a” is added instead when the pathway is ambiguous or unknown.

One of the goals of the Global Eradication Initiative (see page 8150) is to reach a stage where all wild poliovirus transmission is terminated and vaccination can be discontinued. “Emergence of VDPVs,” “failure to vaccinate,” and “vaccine failure” discussed below have been the three major reasons for the delay in achieving eradication of poliomyelitis. Providing that enough money and effort can be mobilized, current

vaccines and vaccine strategies are probably sufficient to enable immediate solutions for “vaccine failure” and “failure to vaccinate.” Promising alternatives applying experience with adjuvants and better applicators have revived the possibility of using techniques explored in the 1950s such as fractional subdermal doses of inactivated virus, IPV [16, 119–121].

Vaccine-derived viruses consistently emerge as a consequence of the inherent genetic instability of poliovirus [122]. Moreover, many of the first sites that mutate restore replicative fitness, reverse attenuation of neurovirulence, and restrictions on growth at elevated temperatures. cVDPVs behave like wild polio [106, 112]. These cVDPVs clearly present a serious health threat [114]. The minimal amino acid changes in neutralizing antigenic sites that occur during person-to-person transmission of cVDPVs [3, 4, 123] allow rapid control through OPV immunization campaigns [106, 114, 124]. In contrast, iVDPV infections have not always been curable [63, 125], and the numbers and identities of anonymous persistent secretors are unknown [109]. The problem of reemergence of poliomyelitis through cVDPVs and especially iVDPVs requires a coordinated global program to discontinue the use of OPV with substitution of alternative vaccination strategies to prevent the appearance of large cohorts of unimmunized individuals during the period when OPV or cVDPVs may still circulate and iVDPV and aVDPV infections persist [116, 122, 126–128]. In fact, eradication should be redefined to include the elimination of both wild and vaccine-derived viruses [122].

OPV strains, like their wild counterparts, readily recombine the noncapsid encoding portions of their genomes with other polioviruses and related non-polio enteroviruses at very early stages in emergence. Evidence for this comes from analysis of poliovirus RNA from vaccine-associated paralytic poliomyelitis cases, VAPP, where, for example, >50% of polioviruses isolates had recombinant genomes [81, 129–132]. Supporting this is evidence from environmental surveillance where vaccine viruses with minimal divergence (0.5–1%) in their VP1 sequences had already recombined with polio and non-polio enteroviral genomes in regions encoding nonstructural proteins [133]. The ability to simultaneously reverse multiple mutations by recombination could foil efforts to

develop improved oral vaccines. Introducing mutations that decrease the chance for reversal by single nucleotide replacement such as incorporation of polymerases with improved fidelity or a total redesign of the genome of each serotype based on rare codon usage [3] could be bypassed by recombination.

Finally, the general consensus is that selective pressure or a higher mutation rate due to local sequence or secondary structure leads to a higher frequency of mutations at certain “hot spots” [93]. However after reviewing the pattern of changes in substitution frequencies throughout the genome it may be more accurate to think of the real frequency of substitutions as that observed in the so-called hyper variable region of the 5'UTR which may be under minimal selective pressure, and consider all other regions as “cold or colder spots” with lower observed rates of substitution derived from negative selection driven by the requirement for viability.

Virion assembly, maturation, and release in picornaviral infections have been reviewed [25]. The ratio of viral particles to infection particles ranges between 30:1 and 1,000:1. Many of the viral particles are noninfectious due to lethal mutations in their genomic RNA and/or incomplete maturation.

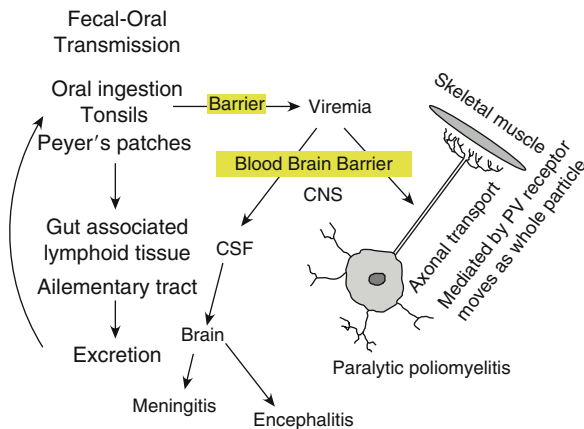
### **Poliovirus Infections at the Level of the Individual Host**

The incidence of poliovirus infections is significantly higher in summer and autumn in temperate zones, becoming less seasonal as the environment becomes more tropical (reviewed in [3]). Improved sanitation and vaccination have reduced natural endemic infections in the very young and together with incomplete vaccine coverage has led to an increasing number of infections in older individuals.

There are two major routes of host-to-host transmission. The most common and most efficient is fecal-oral, followed by oral-oral transmission as Dowdle et al. described for the fate of poliovirus in the environment and their review of the infectious dose for transmission in humans [134]. It has been postulated that there has been a shift from the former to the latter route, as the level of community hygiene improved [56]. The infective dose after ingestion of Sabin vaccine strains is approximately 100-fold higher

than that for wild poliovirus, 1000 CCID<sub>50</sub> compared to 10 CCID<sub>50</sub>, respectively [3, 134, 135]. Nerve damage in the lower spinal cord results in paralysis of the lower limbs (spinal poliomyelitis), whereas damage in the upper spinal cord and medulla may result in bulbar poliomyelitis and paralysis of breathing [72]. The percent of infections ending in paralytic poliomyelitis is further reduced in highly immunized populations. This ratio of asymptomatic cases to paralytic cases has implications for surveillance strategies (see page 8154) based on investigation of all AFP cases. Between these two extremes falls the “minor disease” [72, 136], approximately 5% of infections with wild polio that result in abortive poliomyelitis with fever, fatigue headache, sore throat, and/or vomiting, and another 1–2% result in non-paralytic poliomyelitis with aseptic meningitis, pain, and muscle spasms. The incubation period is between 7 and 14 days but ranges between 2 and 35 days [72]. Virus can be recovered from the throat, blood, and feces by 3–5 days. It was initially thought that viremia was infrequent, but this was based on observations in patients with paralytic poliomyelitis who most likely already had high circulating titers of neutralizing antibodies [72, 137]. However when observations were made early after exposure, for example, in contacts of cases, a high frequency of viremia was demonstrated, implying that the viremia might play a vital role in the development of paralytic poliomyelitis [136, 138]. This was strengthened by concurrent experiments that demonstrated a protective effect against CNS lesions by antiserum in experimentally infected primates. The genetic basis for neurovirulence of poliovirus isolates is addressed below on page 8146.

Paralytic poliomyelitis, encephalitis, and aseptic meningitis occur after a biphasic infection where viremia in some systemic infections is followed by infection of the CNS [136, 138] (Fig. 5). Studies of virus in the CNS and stools in VAPP patients suggested that the virus that invades the CNS was randomly selected [95]. Acute flaccid paralysis (AFP) is a direct result of destructive replication in motor neurons followed by atrophy of de-eneruated muscles. Skeletal muscles are affected when nerves in the anterior horn of the spinal cord are infected and bulbar paralysis occurs when cranial nerves are infected [4, 139]. The maximum effect on muscles occurs within a few days after the start of symptoms. Muscle recovery can occur when



**Polio and Its Epidemiology. Figure 5**

Poliovirus infections. Poliovirus is transmitted from host-to-host by a fecal–oral and to a lesser extent oral–oral routes of transmission. Virus first infects cells in the tonsils, Peyer’s patches, and gut-associated lymphoid tissues and viral progeny are excreted in feces. This phase is followed by a systemic infection during which there is viremia for a short period of time. In some individuals virus crosses the blood–brain barrier by entering the CSF, by axonal transport along nerve cells, and possibly from infected white blood cells that enter the brain. These individuals may develop meningitis, encephalitis, or paralytic poliomyelitis. Destructive viral replication in nerves of the anterior horn of the spinal cord may lead to irreversible atrophy of de-nerved muscles while bulbar paralysis occurs when cranial nerves are infected. Most (>90%) infections of naïve individuals even with the most neurovirulent strains are asymptomatic, 5% result in meningitis, encephalitis, and/or transient paralysis. Only 0.1–1% of the infections will result in permanent paralytic poliomyelitis or death

infection only results in temporary loss of nerve function. Residual paralysis may last from months to the life of the infected individual [72].

Poliovirus infections are not the only cause of AFP. Non-polio AFP occurs with an incidence of 1 per 100,000 children (see page 8153 for the implication this has for surveillance). Guidelines that help epidemiological investigators distinguish AFP caused by polio from AFP caused by other causes are reviewed in Sutter et al. [4]. Final diagnosis requires laboratory confirmation of a poliovirus infection.

Poliovirus infections start as a local infection of cells in the tonsil, intestinal M cells, Peyer’s patch of the ileum, and the mesenteric lymph nodes [3, 72]. This replication in the gut results in the excretion of poliovirus during defecation by all individuals with asymptomatic as well as symptomatic infections and is the basis for fecal–oral transmission. It also provides the rationale for supplementary environmental sewage surveillance (see page 8154) for poliovirus infections. A review of publications between 1935 and 1995 on excretion of polioviruses by Alexander and associates [140] indicated that in most infections of naïve children, wild polioviruses were excreted for 3–4 weeks with a mean rate of 45% at 28 days, and 25% of the cases were still excreting during the sixth week. In contrast, fewer than 20% excreted vaccine strains after 5 weeks. Excretion of polio ranged from a few days to several months [141]. The highest probability of detecting poliovirus positive stool samples was reported to be at 14 days after the onset of paralysis [140] and is the basis of stool sample collection for diagnosis of polio AFP surveillance (see page 8153). The disappearance of poliovirus from sewage samples and from stool samples of immunized children within 6–8 weeks after an immunization campaign [142] or after transition to exclusive immunization with IPV [143] provides additional confirmation for the short duration of excretion. Persistent poliovirus infections are the exception and will be discussed in more detail below. Interestingly, more than one evolutionarily linked lineage of the same serotype may co-circulate in the gut of such persistently infected individuals [79, 104, 107, 109].

Excretion and the duration of excretion are dependent on host factors and on vaccination history of the infected individual. Immunization history may start with passive immunization from maternal antibodies. However, maternal antibodies have an estimated half-life of approximately 1 month [144]. Based on a comparison between titers in cord blood and at 6 weeks, the half-lives for maternal neutralizing antibodies against type 1, 2, and 3 polio were 30.1 days, 29.2 days, 34.6 days, respectively [119]. Immunization history obviously also includes polio vaccinations and natural exposure to endemically circulating wild poliovirus and waning immunity in aging cohorts.

Skeletal muscle injury, including injury caused from intramuscular injections, increases the likelihood

of poliomyelitis in children infected with wild or vaccine poliovirus. Mouse model studies have suggested that in this provocative poliomyelitis, the muscle injury facilitates viral entry to nerve axons and subsequent damage to the motor neurons in the spinal cord [145].

Some individuals who had poliomyelitis develop new muscle pains, hypoventilation, new or increased weakness or fatigue and paralysis decades later after a period of relative stability. This reappearance of polio-related symptoms is referred to as postpolio syndrome. There is a large body of literature relating to postpolio syndrome that will be left up to the reader to pursue. Suggested starting points include the websites of the Post-Polio Health International ([www.post-polio.org](http://www.post-polio.org)), the Mayo Clinic ([www.MayoClinic.com](http://www.MayoClinic.com)), a 1992 paper on the “Epidemiology of the post-polio syndrome” by Ramlow et al. [146], and a 2010 review on the pathophysiology and management of postpolio syndrome by Gonzalez et al. [147]. There is still a debate whether persistent poliovirus or mutated poliovirus contribute to the development of postpolio syndrome [147]. The risk factors include the extent of permanent residual impairment after recovery from the poliovirus infection, an increased recovery after AFP possibly related to the extra stress on compensatory neural pathways and overuse of weakened muscles, the age of onset of the initial illness, and physical activity performed to the point of exhaustion.

Natural infections with poliovirus stimulate both humoral and cell-mediated immunity (see [149, 150] and reviews [4, 148]). Neutralizing antibodies appear in exposed individuals around the time that paralytic symptoms become evident in the few individuals who develop symptomatic infections [72]. Neutralizing IgG and IgM antibodies are also induced in response to immunization with inactivated polio vaccine. The neutralizing antibodies induced after exposure to live or inactivated poliovirus prevent disease by blocking virus spread to motor neurons of the central nervous system [3]. Once seroconversion occurs after vaccination, individuals are protected from disease for life, although circulating antibody titers may wane late in life and may drop below protective levels against one or more serotype in some individuals.

The epitopes on vaccine-derived and wild poliovirus strains that induce neutralizing antibodies may differ from those on vaccine strains. Neutralizing

antibody titers  $\geq 1:8$  against each of the three Sabin OPV serotypes are considered protective; however higher titers may be needed to compensate for the relatively lower antigenicity of wild and vaccine-derived strains [151]. For example, the highest serum neutralizing antibody titers were recorded from individuals immunized exclusively with OPV or IPV when the live challenge virus was the same as that used in vaccination, slightly lower for the respective heterologous strain, and significantly lower for wild and vaccine-derived strains. Serum from some individuals who had titers of  $>1:50$  against Sabin vaccine strains had titers of  $<1:8$  against some wild or vaccine-derived of at least one serotype, suggesting that titers of 1:64, 1:32, and 1:16 against Sabin serotypes 1, 2, and 3, respectively, might be more appropriate to ensure minimal protective coverage [151].

Primary infection in the intestinal tract by wild poliovirus or live attenuated polio vaccine induces secretory IgA antibodies in addition to IgM and IgG antibodies. One of the rationales for the use of live attenuated polio vaccine was that while disease would be prevented by humoral antibody production stimulated by either OPV or IPV, the extent of infection or reinfection and shedding would also be reduced by induction of secretory IgA antibodies by active infection of intestinal cells with live vaccine in mimicry of the natural route of infection [148, 152–155]. IgA induced in the gut plays an important role in terminating primary infection in the intestinal tract and the tonsils [3, 152] affecting both fecal–oral and oral–oral transmission. In practice IPV also induce some intestinal immunity although less than OPV and the duration of excretion in individuals immunized with IPV appears to be longer [23, 56, 156–158].

There is some indication that the duration and possibly memory of intestinal protection is relatively short and the time for clearance of virus relatively longer than the 3–6 or 7–14 days incubation period of the minor and paralytic diseases [152, 156, 159]. Complete blockage of replication in the intestines may occur in only 25–40% of fully immunized children [158, 159]. The rapid decline in intestinal immunity means that polio can establish transient infections even in persons with adequate humoral immunity and circulate silently in that community. Lower efficiency of oral vaccines under certain conditions further complicates efforts to break

chains of poliovirus transmission. Passive immunization with maternal antibodies, which has a short half-life, also affects oral vaccine efficacy (see page 8143).

It is not clear what role cell-mediated immunity may play in the control of polioviral infections. Cell-mediated immune responses were observed early after wild poliovirus infections by the macrophage migration inhibition (MIF) technique but were not observed a later time [160], whereas intradermal administration of subfractional doses caused induration and erythema of 3 mm diameter or above, in 14 of 18 vaccinees that indicated a cell-mediated immune response [161]. In addition, at least in a mouse model, all three serotypes stimulated cross-reactive and serotype-specific T helper cell responses detected by both in vitro proliferation and interleukin (IL)-2/IL-4 production [162].

How can poliovirus infections be prevented? The main tools in the global eradication of poliomyelitis have been the introduction of universal vaccination (vaccine) and improvements in hygiene. The primary goal of routine immunization is to protect the individual [163]. The secondary goal is to immunize a high enough proportion of the population so that the entire population will become protected. As eradication approaches completion, it is becoming more and more apparent that additional approaches will need to be employed in parallel with and perhaps instead of vaccination to extinguish the last pockets of endemic person-to-person transmission and persistent infections. All of these approaches will be necessary to prevent and control reemergence of polio after eradication. For more information, the readers are referred to excellent reviews on inactivated poliovirus vaccine by Plotkin and Vidor [164] and live oral poliovirus vaccines by Sutter et al. [4]. Sources for early history can be found in *A History of Poliomyelitis* by Paul [165] and *Polio Vaccine: The First 50 Years and Beyond* edited by E. Griffiths et al. [166].

The road to the development of effective vaccines against poliovirus was long and paralleled the growing understanding and ability to manipulate viral infections in the laboratory. Most of the important early milestones were listed in the last paragraph of the introduction and in Fig. 1. Mass vaccination trials and studies involving millions of vaccinees played an early and important part in acceptance of universal polio vaccination as a means for fighting

poliomyelitis [19]. It must be stressed that problems and other difficulties during this progression stimulated numerous basic and epidemiological research studies that have resulted in improvements culminating in the current safe high-potency oral and inactivated vaccines that have reduced the number of annual paralytic poliomyelitis cases from >350,000 per year in 1988 to approximately 1,500 in the last few years. Difficulties in reducing this further are discussed below on page 8166. Criteria for quality control for production of polio vaccines introduced by the WHO in 1962 have been updated in relation to newly acquired knowledge about the epidemiology of poliovirus and polio vaccine. One of the major risks associated with the use of live vaccine is that progenies of the vaccine readily accumulate mutations some of which may reverse attenuation. The highest risk for vaccine-associated paralytic paralysis, VAPP, comes from Sabin 3, the vaccine serotype that also has the highest variability across production lots [3]. However the risk of OPV-associated polio is less than 0.3 per million doses [22] with the risk highest in naive children receiving their first dose [81]. The risk (see page 8156) of not using oral vaccine for global eradication compared to its use at the current stage in the Global Poliovirus Eradication Initiative remains overwhelmingly in favor of its use [1].

Three incidents nearly derailed early efforts to develop and employ effective vaccines. The most glaring of these, primarily from the point of views of negative publicity for use of polio vaccines, was the “Cutter Incident” in 1955 where wild poliovirus was inadequately inactivated probably because of failure to remove clumps that may have sequestered and protected infective vaccine virus, and a nonlinear tailing-off of inactivation at low titers [22, 167, 168]. Altogether more than 400,000 children were inoculated with an inadequately inactivated vaccine batch produced by the cutter vaccine production facility which resulted in 94 cases of poliomyelitis among primary vaccinees, 126 cases in family contacts, and another 40 cases among community contacts and 10 deaths. The publicity caused great concern throughout the world until the cause was discovered and corrective measures applied. The second, apparent failure of early vaccines to protect against subsequent infection and paralytic disease due to an initial lack of awareness in the 1950s that there were three non-cross-reacting serotypes of

polio has already been mentioned. The third problem, the contamination of live polio vaccine with SV40 virus, a simian virus, continues to raise concerns about long-term effects from human zoonotic infection with this virus that was shown to cause cancer in mice [22, 169–171]. The SV40 was inadvertently introduced through the use of SV40-infected simian cell cultures in some early vaccine production batches. So far there is little evidence for any contribution to the incidence of tumors in the humans who received SV40.

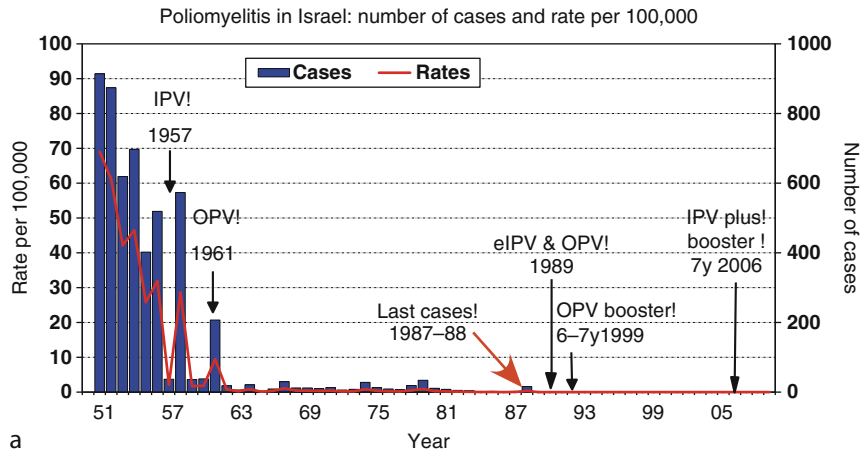
The many vaccination formulation and vaccination schedules that have been employed during the effort to eradicate polio and the rationale for their use have been reviewed in depth [4, 164]. Changes in schedules and formulations mean that in any one region different cohorts in the total population will have received different vaccine formulations and immunization schedules. This complicates determining duration of protection and interpretation of events. The evolution of vaccination policy in Israel [172–174], a graph showing the history of poliomyelitis in Israel (Fig. 6), and the two disagreeing discussions that were published within the same report on the underlying causes that enabled the last outbreak in Israel in 1988 [175] are a good example of this difficulty.

Isolation of poliovirus with attenuated neurovirulence was a prerequisite for the development of oral polio vaccines (OPV; see reviews [3, 4, 166]). Vaccine candidates were either derivatives of neurovirulent or even highly neurovirulent (e.g., Sabin 3) isolates selected for attenuation after passage in primates, primate cell cultures, and/or non-primate cell cultures or starting from isolates with low neurovirulence (e.g., Sabin 2). Neurovirulence refers to the ability of an isolate to cause an infection adversely affecting functions of the CNS, keeping in mind that for any neurovirulent isolate, only 5% of infections cause transitory adverse CNS effects and less than 1% cause permanent paralytic poliomyelitis. The total number of nucleotide differences between vaccine strains and their respective parental strains was found to be 57 nucleotides and 21 amino acids for serotype 1 [176–178], 2 nucleotide differences and 1 amino acid difference for serotype 2 [179, 180], and 10 nucleotide differences and 3 amino acid differences for serotype 3 [181, 182]. Sequence analysis coupled with genetic manipulation

has allowed investigators to pinpoint which of these nucleotide differences between vaccine candidates and vaccine strains account for the loss of neurovirulence (Fig. 3) [93, 183]. “Quantitative determination of the contributions of each substitution is complicated by several factors: (a) The role of minor determinants of attenuation is difficult to measure, (b) some substitutions have pleiotropic effects on phenotype, (c) some Sabin strain phenotypes require a combination of substitutions, (d) second-site mutations can suppress the attenuated phenotype in various ways, and (e) the outcome of experimental neurovirulence tests may vary with the choice of experimental animals (monkeys versus transgenic mice) or the route of injection (intraspinal versus intracerebral)” [3]. The propensity of vaccine to evolve and revert to neurovirulent phenotype is discussed throughout the current review.

The safety and effectiveness of live attenuated polio vaccine strains in preventing poliomyelitis was very clearly demonstrated in large clinical studies involving millions of children in the 1950s [19]. A number of factors including vaccination schedules, the presence of maternal antibodies, hygiene, and nutritional status of the individual influence the efficiency of induction of seroconversion by OPV strains. Early studies showed that viral interference between strains in the trivalent vaccine and from concurrent infections with non-polio enteroviruses also influences vaccination outcome [19]. Multiple doses of OPV are recommended to ensure seroconversion rather than to boost waning immunity [184]. The number of OPV doses that is needed to reach 90–95% seroconversion rates in naïve children is not the same for all populations. For example, three doses will seroconvert 90–95% of naïve children in developing countries, whereas in certain regions within developing countries such as India, the same three doses will only seroconvert a maximum of 60% of vaccinees [184]. Supplemental immunization activities (SIAs) employed sometimes more than once a year are needed to ensure adequate primary vaccination coverage and to fight endemic circulation of wild poliovirus or reintroductions of wild poliovirus. In SIAs, all children in national or subnational regions are immunized in national immunization day (NID) and/or subnational immunization day (SNID) campaigns, respectively, with a dose of OPV irrespective of vaccine history. The costs of the additional doses





Polio vaccine immunization schedules in Israel 1957–2010

- 1957–1960 exclusive use of IPV (mainly local production)
  - varied number of doss (2 to 4)
  - varied quantity of antigen per dose
  - initially administered in campaigns then by routine vaccination schedule!
  - estimated coverage 75% to 90%
- 1961–1963 exclusive use of OPV
  - coverage 95%
  - 1961 mOPV1
  - 1962 add mOPV2 and mOPV3
  - 1963 tOPV 3 dose schedule (2, 6 and 12 months)
- 1964–1978 4 doses by (2, 4, 6, 24 mo) coverage 81% to 91%
- 1979–1981 same as in 1964–78 plus a supplemental mOPV1 once a year for ages 0–2 yrs
- 1982–1987 three programs
  - a. 4 OPV doses in 12/14 health sub-districts
  - b. 4 OPV doses as in a. plus one dose mOPV1 in selected groups at risk
  - c. exclusive eIPV 3 doses in 2 of 14 health sub-districts at least one dose in combination with DTP.
- 1989–2005 combined eIPV/OPV
  - 2 doses of eIPV (2 and 4 mo)
  - 2 OPV doses (4 and 6 mo)
  - simultaneous IPV plus OPV at 12–16 mo)
  - 1990 IPV booster added at 6–7 yr
- 2006–2010 exclusive use of eIPV
  - 4 doses of eIPV (2, 4, 6, 12 mo) plus booster 7 yrs)
- 1957–2010 Immigration of families with children who were vaccinated by different vaccination schedules used at their countries of origin.

b

**Polio and Its Epidemiology. Figure 6**

*Prevention of poliomyelitis through universal vaccination and evolving vaccination strategies as illustrated by the history of cases and vaccination schedules in Israel between 1957 and 2010.* Figure (a) represents the annual number of cases (blue bars) of laboratory confirmed poliomyelitis cases and the rate per 100,000 children (red line) between 1951 and 2010. The red arrow indicates the last cases of poliomyelitis that occurred during an outbreak in 1987–1988. Israel has been poliomyelitis-free since 1989. Black arrows indicate major changes in vaccination policy. Previous attack rates of 14.2 and 146.9 per 100,000 in 1949 and 1950, respectively, signaled the transition from endemic to epidemic epidemiology of poliomyelitis in Israel. A full list of vaccination schedules is indicated in (b) (Data presented in (a) was supplied with permission by the Israel Center for Disease Control. The vaccination schedules were taken from Swartz TA. The Epidemiology of Polio in Israel A Historical Perspective. Tel Aviv: Dyonon Pub. Ltd.; 2008 [172])

needed to raise seroconversion rates to above 90%, significantly raise the cost for effective immunization with OPV and require the coordination and use of many paid and voluntary staff. In fact, in the end it may actually be easier to immunize three times with IPV (even at current costs) than with the additional number of doses of OPV especially when access to populations is difficult and environmental conditions challenge maintenance of viability of the live vaccine. This counters both the lower cost and difficulty of administration rationales for using OPV instead of IPV. Mass immunization campaigns have rapidly boosted herd immunity [3].

The take of OPV is negatively influenced by the presence of maternal antibodies. Nonetheless, when infants are fed OPV at birth, 30–60% excrete virus, 20–40% of infants seroconvert, and the subsequent take of OPV is better when a birth dose is given (reviewed in [184, 185]).

Vaccination formulation must also take into account differences in the efficacy of induction of intestinal immunity by the different vaccine serotypes [155]. For example, type two was 100% effective with two doses, whereas types 1 and 3 needed more than three doses. Since the elimination of wild type 2 in 1999 [186], and the significant decrease in the number of endemic regions where wild type 1 and 3 co-circulate, SIAs have increasingly turned to the use of monovalent and divalent OPV. Monovalent OPV vaccines improve seroconversion rates compared with tOPV [187]. However routine immunization still requires the use of tOPV to prevent the accumulation of large cohorts of individuals who are naïve to type 2 poliovirus and who could serve as a reservoir for transmission of neurovirulent type 2 VDPV as has occurred in Nigeria [106]. New guidelines for the use of mOPV1, mOPV2, and dOVP1+3 have been recently issued [188].

The use of inactivated poliovirus is an alternative approach to vaccination against polio (reviewed in [164]). Salk developed an inactivated polio vaccine, IPV, using neurovirulent strains of the three serotypes of poliovirus. IPV was successfully tested by a placebo-controlled trial in over 400,000 children and in unblinded observations on another 1,000,000 children before certification for use in the mid-1950s [16, 17, 189]. A relatively higher difficulty in production, greater production costs, higher difficulty in

administration, and the initial belief that only live vaccine would efficiently evoke intestinal immunity led to the choice of OPV for most routine national vaccination programs [122]. Countries are currently switching to vaccination with IPV in combination with OPV or more often in place of OPV because of its relative safety record (no VAPP cases), improvements in manufacture that have increased effectiveness and reduced the cost difference between a dose of OPV and IPV, and the paradoxical success of OPV in reducing poliomyelitis as an epidemic disease in most countries [164]. Additional motivation has come from the increasing awareness that fully neurovirulent vaccine-derived polioviruses behave like wild polioviruses [133, 190, 191] that must be eliminated and prevented from emerging in order to attain final success for poliomyelitis eradication.

Early studies on genetic and antigenic variation such as a study of Sauket strains, the type 3 used in production of IPV [192] were instrumental in the establishment of rigorous standards for seed stocks for vaccine production. The original IPV formulation has since been improved. This enhanced IPV, eIPV, has a higher immunogenicity and protective efficacy than IPV [157]. A number of factors contributed to this improvement. These included new production protocols, new tissue culture techniques including a microcarrier-based technology, and a more optimal balanced formulation of the three serotypes. It can be administered alone or can be combined with other vaccines such as DTP. The immunogenicity of eIPV was at least as good as that of OPV and there was good long-term immunity [157]. Subdermal administration of fractional doses of IPV was one of the approaches tried in the early 1950s [16]. Subsequently seroconversion rates from fractional doses were shown to be adequate but somewhat lower than for full dose intramuscular injections fractional doses [119, 193]. In another trial, similar seroconversion rates were observed but there were lower median titers in those receiving fractional doses [120]. Fractional doses effectively boost titers in previously immunized individuals [194]; however there is no long-term information on the rate of waning immunity in individuals treated with these fractional doses. Large non-inferiority studies testing subdermal administration of fractional doses of IPV using needle-free devices such as recently by

Mohammed et al. [120] and Resik et al. [119] offer one quite promising practical solution for realizing cessation of use of live OPV with affordable alternative vaccines as recommended by the Advisory Committee on Polio Eradication in 2004 [195].

Antiviral drugs offer a promising complimentary or alternative approach to the use of vaccine to control poliovirus infections especially for persistent infections in immunodeficient individuals, during the final stages of eradication, and for post-eradication reemergence [122]. Presumably these drugs may also work to control severe infections by non-polio enteroviruses or have been chosen because they have been shown to do so. Drugs with unique virus-specific targets such as capsid proteins, the hydrophobic pocket, the RNA-primed RNA polymerase, protease inhibitors, protein 3A inhibitors, nucleoside analogs, proteinase 2c inhibitors, and compounds with unknown mechanisms of action have been reviewed [196]. There is still a long way to go to find truly effective universal anti-polio or anti-enteroviral drugs, thus only a few examples will be provided.

Pocket factor drugs such as WIN 51711 [74], isoflavones [61], pleconaryl [197], and capsid inhibitor V-037 [198] prevent viral entry by interfering with receptor binding or by preventing conformational changes needed for viral capsid uncoating. One of the difficulties in developing pocket factor drugs comes from the quasispecies nature of enteroviral infections, where mutants may rapidly emerge [61, 63] or there may be viral isolates already present in the quasispecies that have mutations in the capsid that may either render the isolate resistant or even dependent on the drug for growth. Furthermore these resistance mutations may not even have to be at the drug binding site (see, e.g., [61]).

Ribavirin is a drug that normally interferes with mRNA capping. While enterovirus mRNA is uncapped, the polio polymerase can incorporate ribavirin into both negative and positive-strand progeny RNA molecules increasing mutagenesis above the catastrophe limit causing a decrease in the reproductive capacity of the viruses [66, 94] (discussed above on page 8138).

Passive immunization has also been tested as a means of preventing polio. Administration of immunoglobulin shortly after exposure to polio may reduce the incidence or severity of paralytic disease although

its general use is not practical due to the short time during which it is effective [199]. Intravenous preparations of immunoglobulin prepared from human populations exposed to enteroviral infections have however helped to decrease chronic meningoencephalitis infections by these enteroviruses in agammaglobulinemic patients [200]. Regular intravenous treatment may help prevent poliomyelitis in immunodeficient individuals but may not prevent virus replication and shedding [62]. Passive immunization with immunoglobulin or human milk rich in anti-polio IgA together with another antiviral pleconaryl may have helped to resolve at least one persistent poliovirus infection [125]. However, efforts to cure another persistent poliovirus infection with human milk and ribavirin, or other antiviral treatments, did not prove successful [63] and this individual has continued to excrete highly diverged vaccine-derived poliovirus for more than 20 years [62]. Anecdotally, shedding of intestinal mucosa associated with a superinfection with *Shigella sonnei* may have helped to cure another persistent excretor [62].

### Poliovirus Infections in Populations

Poliovirus infections in populations have been the subject of many reviews over the years. The older reviews are still of interest not only because of the information they review but because they also provide a picture of policies and knowledge available at the time. The following paragraphs will concentrate on those aspects of poliovirus infections in populations that impact the most on the endgame strategy of poliomyelitis eradication. The discussion will start with a brief overview of the changing nature of the epidemiology of poliovirus infections. This will be followed by a description of the Global Polio Eradication Initiative and will end with a discussion of the three main problems that have led to a delay in its realization, namely, “failure to vaccinate,” “vaccine failure,” and “vaccine-derived polioviruses.” When reading this section which will highlight some of the current problems and their solutions, the reader must keep in mind the overwhelming success of the Global Polio Eradication Initiative to date: a major reduction in the number of endemic countries where polio is still transmitted from 126 to 4; a decrease in the number of annual cases by >99% that prevented

life-long paralysis in more than five million children between 1988 and 2005; eradication of one wild poliovirus serotype in 1999; and elimination of the majority of wild lineages throughout the world.

The nature of poliovirus infections in populations has gone through a number of phases. Before the appearance of outbreaks of poliomyelitis starting in the nineteenth century, poliovirus circulated endemically. Infections occurred in the very young, and conferred lifelong immunity against reinfections with the same serotype. The epidemics became more frequent, grew in size, and infections included older children and adults who were not naturally immunized. Vaccination has drastically reduced the number of people who have been exposed to natural infection with wild polioviral strains. There also appears to be a shift in person-to-person transmission routes. Oral–oral transmission has increased in importance while fecal–oral transmission decreased as a result of improved hygiene [56]. Control of poliomyelitis requires breaking all chains of wild poliovirus transmission by immunizing all children with three doses of polio vaccine or at least enough children so that herd immunity protects the remaining population. The percent of the population that needs to be immunized for establishment of herd immunity against wild poliovirus is >85% for developed countries and >90–95% in tropical developing countries. In 2010, there were still three groups of countries: four “endemic countries,” Afghanistan, India, Nigeria, and Pakistan where the transmission of wild viruses has not yet been completely halted, “polio-free countries” where vaccination has broken all endemic chains of wild poliovirus circulation and there have been no cases other than VAPP within the last 3 or more years, and “importation countries” that were formerly polio-free, but where there are poliomyelitis cases caused by wild poliovirus imported from one of the “endemic countries” and where there may be local transmission of the imported wild poliovirus. Between 2002 and 2006, there were 26 importation countries, 7 with viruses imported from India and 19 with viruses imported from Nigeria [4]. There were 21 importation countries in 2009 and 13 in 2010. These included ongoing outbreaks in Tajikistan and the Russian Federation and apparently expanding outbreaks in Angola and the DRC. The former represents the first cases from wild poliovirus in the European region since regional

eradication was declared in 2006 and the latter could potentially spread to polio-free countries in Africa and other regions. The list of importation countries is updated on a weekly and monthly basis and can be accessed at the web page of the Global Polio Eradication Initiative, [www.polioeradication.org](http://www.polioeradication.org). Transmission routes are dependent on population movements. One or more outbreak founders may be introduced by infected persons coming from an endemic or importation country or by returning travelers from such countries. One event with very high risk for spreading poliovirus from one country to another is the Hajj in Saudi Arabia. To counter this threat, it is now mandatory for foreign pilgrims coming on Hajj and Umrah to be vaccinated for communicable diseases including polio, especially pilgrims with young children arriving from countries with polio cases. The children must have received a dose of OPV 6 weeks before their arrival and another upon arrival.

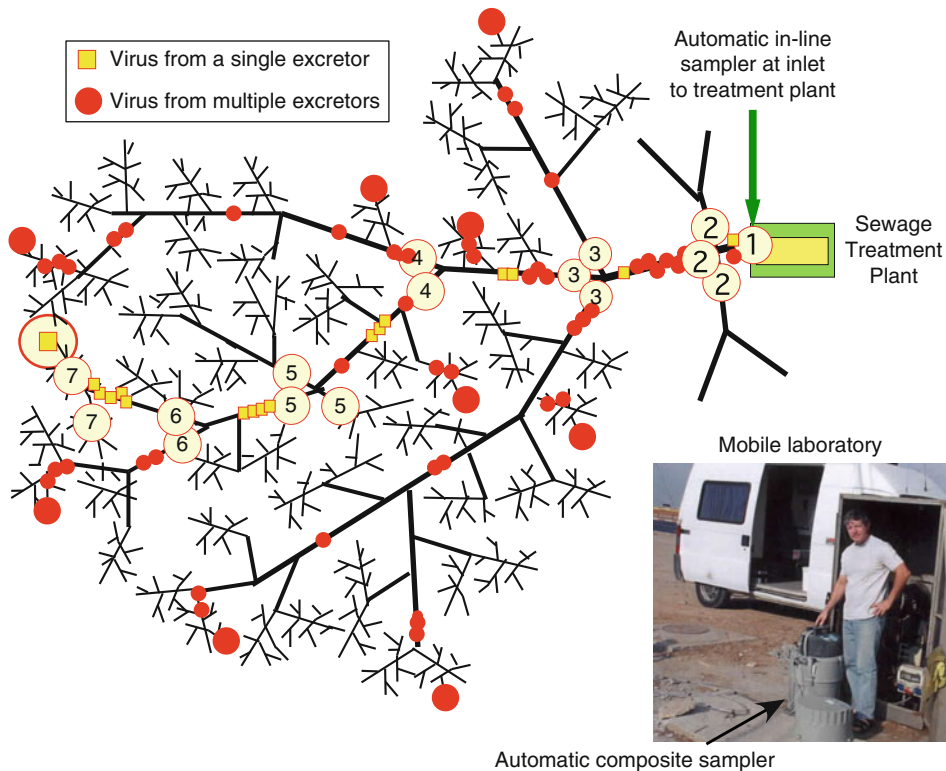
Smallpox was eliminated as a circulating human pathogen in 1977 after an 11-year extensive vaccination and surveillance program [201]. Only two sources of smallpox virus have been reserved for research purposes, one in the United States and one in Russia. Final eradication will be achieved when these last two remaining, contained sources of smallpox virus are finally destroyed. Proposing a similar approach in 1988, the World Health Assembly set a goal of eradicating poliomyelitis by the year 2000 (resolution WHA41.28). A group of experts at the global, regional, and country level set the criteria and conduct the process of certification of eradication [202]. These experts must be independent from the vaccine administration and polio laboratories. Global polio eradication, first targeted for completion by 2000, was limited to the eradication of all wild polioviruses with the caveat that “the occurrence of clinical cases of poliomyelitis caused by other enteroviruses, including attenuated polio vaccine viruses, does not invalidate the achievement of wild poliovirus eradication” (Report of the first meeting of the Global Commission for the Certification of the Eradication of Poliomyelitis. Geneva: World Health Organization; 1995. WHO document WHO/EPI/GEN/95.6). OPV vaccination would cease within a few years after eradication of poliomyelitis from wild polio and industrialized nations would save not only the large sums of money needed for the

maintenance and rehabilitation of individuals with paralytic poliomyelitis but the costs of vaccine and vaccine administration as well [203–206].

In order to easily eradicate an infectious agent, (a) the agent should replicate in a single host with no intermediate vector, alternative reservoir species, or carrier state, (b) vaccines and/or anti-infectious agents must be available to break chains of transmission whether directly between susceptible organisms or after exposure to the infectious agent in the environment, (c) if transmission involves environmental exposure, there must be a finite and relatively short survival time for the infectious agents in the environment, (d) all or the majority of infections should be clinically apparent with unique symptoms, and (e) there must be an easy and cost-effective surveillance system for detection of the infectious agent, identifying infected individuals, and for determining the efficacy of treatments in individuals and populations [3, 207]. Deviations from some of these requirements for eradication have made eradication of polio more difficult from the start [56]. In particular, most (>95%) poliovirus infections are clinically asymptomatic while symptoms associated with the few infections that result in poliomyelitis are not unique to poliovirus infections and although effective vaccines were available, there is no easy and cost-effective method to determine the effectiveness of vaccination in vaccinated individuals. This makes surveillance and the identification and isolation of infected individuals much more difficult. When the GPEI resolution was passed in 1988 and even as late as 1996, it was stated [208] that there was no indication of chronic excretors; however, persistent infections (see page 8161) do exist and have emerged as one of the difficulties in achieving eradication. Differences between smallpox and polio have made it relatively more difficult to achieve polio eradication. For example, live polio vaccine is made from three temperature-sensitive serotypes of poliovirus because there were three non-cross-reacting poliovirus serotypes, whereas the vaccine for smallpox was a single more stable unrelated bovine virus. This has complicated vaccine formulation and administration. In addition since the vaccine contains live poliovirus there are also a number of more severe safety issues concerning pre- and post-eradication vaccine production compared with the smallpox vaccine.

Specific eradication strategies for polio included (a) high routine immunization with at least three doses of vaccine for all children and an additional birth dose in countries where polio has remained endemic (b) SIAs, either national or subnational immunization campaigns targeting children under 5 years of age (c) surveillance, primarily investigation of all cases of AFP with an increase in the number of supplementary surveillance programs such as sewage surveillance and enterovirus surveillance as the endgame of eradication approaches, and (d) house-to-house mopping-up immunization campaigns to block final chains of wild polio transmission [1, 163, 203]. The WHO requires genomic sequencing of all isolates of potential interest to the Polio Eradication Initiative. An isolate is of interest when results from either standard immunological and/or molecular tests conducted by accredited laboratories using standard methods (see next section) indicate that the isolate has behaved differently than standard Sabin strains of the corresponding serotype.

A Global Polio Laboratory Network, GPLN, was established to monitor poliovirus infections throughout the world using standardized methods, cell lines, reagents, and reporting methods [209–213]. These standard methods (WHO/EPI.CDS/POLIO/90.1) have been reviewed and revised and improved as knowledge about the epidemiology of polio expands and as new analytical methods become available (Fig. 7). This includes even the flow charts or “algorithms” for culturing viruses, identifying polioviruses, and characterizing the serotype (Typic Differentiation) and determining wild, vaccine, or vaccine-derived virus within specific time frames (Intratypic Differentiation). The current fourth version of the *Polio Laboratory Manual* (WHO/IVB/04.10) was adopted in 2004. The three levels of laboratories, National and Sub-national Laboratories, Regional Reference Laboratories, and Global Specialized Laboratories, are certified each year through on-site visits, after accurate testing and timely reporting of a minimum number of relevant assays, and by results from proficiency tests. The requirements for certification, quality assurance, and safety and the responsibilities of the three types of laboratory are spelled out in the *Polio Laboratory Manual* (WHO/IVB/04.10). By the end of 2009, the GPLN consisted of 146 laboratories of which 139 were fully accredited by the WHO, and another 5 in the process of accreditation (WHO 16th



### Polio and Its Epidemiology. Figure 7

*Population-based environmental surveillance for poliovirus.* The figure is a schematic representation of a network of sewage drainage pipes leading to a sewage treatment plant starting from individual homes, schools, or places of work or entertainment (*thinnest lines*) and converging into larger and larger *trunk lines* (*thicker lines*) until entering the treatment plant. There is an in-line automatic sampler at the inlet to the treatment plant (*green arrow*). Portable automatic samplers like those illustrated in the photograph (*black arrow*) can be lowered into sites at branch points to determine which of the branches contain virus detected by downstream sampling sites. *Yellow squares* represent virus excreted by a single infected individual living at the periphery of the system (*large yellow square*). *Large red circles* represent the situation where more than a single individual is infected and the viruses that they excrete are represented by the smaller *red circles*. The incrementally increasing *black numbers in the circles* represent upstream the order in which the portable samplers can be placed at major branch points to approach and determine the location of the single excretor (Adapted from Hovi T et al. [143] and Shulman LM et al. [291])

Informal Consultation Of The Global Polio Laboratory Network, September 2010, Geneva, Switzerland).

The Polio Laboratories are coordinated on a regional basis by Regional Laboratory Coordinators who report to the Global WHO Polio Coordinator in Geneva. Identification tasks such as intratypic differentiation and sequence analysis originally assigned to the more specialized labs are now being certified for use in National and Regional Reference Labs as expertise increases and methods – especially molecular methods are simplified. This trend has been accelerated by the

increasing difficulty and costs of shipping material that may contain live infectious wild polioviruses between the different levels of laboratories and the need for decreasing the time between isolation and final notification of characterization of the poliovirus isolates. Rapid identification is especially critical for eradication efforts in endemic regions and for identifying introductions to polio-free regions from these endemic regions.

The laboratories work in close coordination with epidemiologists and medical staff in the investigation

of all cases of AFP and/or supplementary enterovirus surveillance and with municipal employees and epidemiologists where supplemental environmental surveillance is utilized to screen for poliovirus presence and circulation. In late 2010, a commercially available method for preparing noninfective viral RNA suitable for molecular analysis based on automatic nucleic acid extraction, immobilization, and storage on Flinders Technology Associates (FTA) filter papers was being evaluated to increase safety and drastically reduce costs of shipping material between laboratories (Summary and Recommendations of the 16th Informal Consultation Of The Global Polio Laboratory Network, 2010, Geneva). Using this technology, viable virus could be reconstituted from the RNA after it is transfected into eukaryotic cells.

Intratypic differentiation (determination of the serotype of an isolate) was based on results from one immunological ELISA test [214] and one molecular-based test, either probe hybridization [215], diagnostic RT-PCR [216], RT-PCR and RFLP [217], RealTime-RT-PCR [218], or micro-array-based systems [80]. At the 16th Informal Consultation Of The Global Polio Laboratory Network, 2010, Geneva, the recommendation for ITD testing from ITD-accredited laboratories was modified to one of the following three options: (a) two RealTime RT-PCR procedures, one for ITD and one for detecting vaccine-derived poliovirus and sending all non-Sabin-like viruses or Sabin-like viruses with non-Sabin-like VP1 to higher level labs for full-length sequencing and molecular analysis of VP1, (b) one validated ITD method (ITD, or molecular) and shipment of all viruses to higher level labs for full-length sequencing and molecular analysis of VP1 or (c) on-site full-length VP1 sequencing of all isolates or referral of all virus isolates to higher level labs for full-length sequencing and molecular analysis of VP1. Results are confirmed by sequence analysis of the entire VP1 capsid gene. Molecular analysis of the sequences of the genomic RNA encoding the VP1 capsid gene is in fact the definitive method to determine whether an isolate is a vaccine strain, a VDPV, or wild isolate. Sequence analysis of the VP1, all four capsid genes (e.g., P1), and even the entire genome, infers evolutionary relatedness to other isolates in endemic or external reservoirs. The methodology and results from such analyses that help trace the origin of viruses founding outbreaks have

been clearly presented in reviews by Kew et al. [101] and Sutter et al. [4]. Sequence data is kept in databases maintained by the specialized laboratories of the Polio Laboratory Network, such as the CDC in Atlanta, GA, Pasteur Institute in Paris, and the HTL in Finland. Phylogenetic comparisons of sequences from new isolates, routinely provided by the CDC, indicate evolutionary relationships to previously isolated polioviruses from the same region and trace importations to or from external reservoirs. Important epidemiological information can be obtained from this phylogenetic analysis. For example, a significant gap between a new sequence and all other known sequences indicates a gap in surveillance while an importation implies that there are cases or silent circulation of related viruses in the region containing the reservoir that it is most closely related to. Knowledge obtained about time clocks [39] for nucleotide substitutions (see page 8139) allow investigators to infer whether nucleotide differences between two isolates are consistent with local transmission or represent separate introductions (e.g., see Manor et al. [102]).

The currently recommended standard method for poliomyelitis surveillance is based on the isolation and molecular and serological analysis of all viruses from all cases of acute flaccid paralysis, AFP, to rule-in or rule-out polioviral etiology [219]. The definition of an outbreak varies depending on whether endogenous poliovirus transmission has remained unbroken or the area has been found to be polio-free. In the latter, given the goal of eradication, a single AFP case can be considered to be an outbreak. The previous section describes what is needed for timely high quality testing of all poliovirus isolates from cases and from other sources such as environmental surveillance. Much time, effort, and money has been spent on maintaining and improving lab quality assurance and performance of laboratories in the Global Laboratory Network. However two factors outside of the control of the laboratories strongly influence the final result. The first is sample collection and the second is the conditions under which the sample is stored and shipped to the first processing lab. The most appropriate sample with the highest probability of detecting poliovirus is a 5-g stool sample. For AFP cases, two stool samples (not rectal swabs) should be collected 1–2 days apart within 14 days of onset of paralysis. This is based on

a review of studies measuring the timing of viral excretion (discussed on page 8143) in infected individuals [140] and the fact that detectable viral excretion is sometimes intermittent. Standardized tissue culture conditions using limited passages of poliovirus sensitive L20B, HEp2C, and RD cells provided by Specialized Laboratories of the Global Polio Laboratory Network are used according to standard operating procedures to isolate polioviruses from clinical samples [219]. An amended algorithm for isolating polioviruses designed to reduce the workload and the time between receipt of sample and identification of viruses of interest has been successfully evaluated in a number of National Poliovirus Laboratories (WHO 16th Informal Consultation Of The Global Polio Laboratory Network September 2010, Geneva, Switzerland). Standard typing and intratypic differentiation assays are based on results from serological assays and molecular assays as described above with final characterization based on the full-length sequence of VP1. Additional regions such as the 5'UTR, the entire P1 region encoding all four capsid proteins, the 3D polymerase, or the entire genome may be sequenced for higher resolution and to determine whether and to what extent genomic recombination has occurred.

Molecular data from any polioviral isolates recovered from the stool samples provides information about the serotype of the isolate or isolates, and differentiates between VAPP, persistent VDPV, and circulating VDPV or wild polioviruses. The different time clocks for single nucleotide substitutions [39] and unique recombination patterns are important tools for these analyses. Timely AFP surveillance also provides the necessary critical information about the temporal and geographic distribution of the isolates for efficient and economical infection and outbreak response. The rationale for AFP surveillance is based on the observation that AFP from all non-polioviral causes occurs with an incidence of 1 per 100,000 in children up to the age of 15. When all AFP cases are investigated and the AFP incidence is within the range for non-polio causes, the absence of poliovirus in the stool samples from any AFP case is considered to indicate absence of circulating poliovirus in the region under surveillance. A surveillance area is considered to be wild poliovirus-free when adequate AFP surveillance levels for greater than 3 years indicate absence of

wild poliovirus, and entire WHO-designated regions are considered to be free from endogenous wild polioviruses when all countries within that region are wild poliovirus-free.

Wild poliovirus positive AFP cases in previously polio-free areas or WHO regions can occur. Molecular analysis then reveals the most likely external reservoir from which the virus was transmitted [101]. Two recent examples of country-to-country transmission (see page 8156) which have seriously impacted the eradication initiative are the spread of wild polioviruses to >21 polio-free countries [220] as consequences of the temporary cessation of vaccination in Nigeria starting in 2003 and the spread of wild polio into the European region in 2010 [221] enabled by low vaccine coverage in Tajikistan. Sequence analysis of poliovirus isolates from cases in Mumbai confirmed cessation of local chains of transmission and the reintroduction of viral lineages still circulating in the north of India [222].

Most countries employ AFP surveillance. However not all are able to reach the required incidence of AFP investigations. Some of these countries supplement AFP surveillance with enteroviral surveillance and/or environmental surveillance (Fig. 7). Enteroviral surveillance is the systematic identification of the enterovirus genotypes in all clinical infections in general or more specifically from all cases with associated meningitis and encephalitis, symptoms that may appear more frequently than AFP in patients with poliovirus infections (approximately 5% of poliovirus infections compared to 0.5–1% for AFP). In some countries enteroviral surveillance and/or environmental surveillance are used exclusively.

A number of different sampling techniques have been used to obtain environmental samples including grab sampling during peak sewage usage, trapping with silicates or gauze, and automatic composite sampling of sewage aliquots at given time intervals over a 24-h period (Guidelines for environmental surveillance of poliovirus circulation, WHO/V&B/03.03 [223]). All sample storage and shipment must be at low temperatures (4–8°C) to maintain viability since currently certified tests require an amplification step in tissue culture. The FTA technology trial referred to above may eliminate the need for maintaining low temperatures.

The usefulness of L20B cells to isolate polio isolates in the presence of high titers of non-polio



enteroviruses [222] has already been mentioned. It is still important to characterize the viruses that grow since L20B cells can also support growth of some other human and bovine enteroviruses, as well as less well-characterized viruses [56]. Additional steps involving molecular screening and growth at elevated temperatures has enabled investigators to more easily identify and characterize wild and vaccine-derived polio in the presence of high titers of vaccine viruses [102, 215, 216, 218, 224, 225]. Selective growth of non-vaccine poliovirus at elevated temperatures [226] is based on a relative small decrease of titer for these viruses compared to a much higher reduction in yield for vaccine strains at elevated temperatures. The main molecular determinant responsible for this difference is a single nucleotide change in loop V of the 5'UTR which can revert or be modified by other changes, hence some polioviruses of interest may escape detection and some minimally diverged vaccine virus may be included among the selected isolates. Confidence that polioviruses isolated by environmental surveillance reflect circulating viruses comes from the high sequence homology between environmental samples and isolates from cases [38, 143, 227]. One of the major contributions of environmental surveillance reviewed by Hovi et al. [143] is that it can be used to establish the presence and/or circulation of wild or vaccine-derived polioviruses before the appearance of AFP cases [102, 222, 227–229]. Environmental surveillance has also revealed the presence of presumptive primary vaccinees excreting OPV in Switzerland where vaccination is by exclusive use of IPV [56].

Different methods for analyzing environmental samples are also currently employed in different laboratories [223] since unlike AFP surveillance [230] there is as of yet no single standard method. The probability of detecting poliovirus in environmental samples [229] depends on the duration and amount of poliovirus excreted by one or more infected individuals (see page 8143), the effect of physical and mechanical factors on the dilution and survival of poliovirus in the sewage system (reviewed by Dowdle [134]), and the frequency of collection and laboratory processing of the environmental samples [223]. A model based on these factors [229] predicted that environmental surveillance could outperform AFP surveillance for small outbreaks as well as detect circulation before the appearance of

cases. The location of the sampling site relative to the excretor and the number of excretors (Fig. 7) also determines the probability of detection [143]. In general polioviruses can be quantitatively recovered from the environment [231]. Decreasing this distance between the excretor and the sample collection site is more effective in increasing the probability of detection and less labor intensive than increasing the sampling frequency [109]. Environmental surveillance is resource and labor intensive and may require large capacity high-speed centrifuges that are not commonly present in most National Poliovirus Laboratories [143]. It requires judicious choice of potential target populations, a competent laboratory, a plan for routine surveillance and reporting, and the cooperation of municipal authorities. The WHO has recommended principles for selecting sites, sampling strategies, and interpretation of results [223]. Lengthy periods of poliovirus-free monitoring are needed to confirm that poliovirus transmission has stopped since even the most comprehensive surveillance covers only subgroups of the entire population of potential excretors [232].

Many poliovirus positive environmental samples contain one or at best a few polioviruses of interest indicating that the surveillance is operating at the lower limits of detection. Thus while negative findings cannot rule-out the presence of polioviruses at levels below detection, they gain significance when they are part of a long sequence of negative results from frequent routine surveillance at the given site. A positive finding of a wild poliovirus or a VDPV can trigger a response ranging from public announcements to remind individuals scheduled for routine vaccinations to be vaccinated in time in areas with high vaccine coverage to scheduling NIDS or SNIDS in regions where immunization coverage is below that required for establishing herd immunity. Sequence analysis can differentiate between multiple importations and local circulation when more than one poliovirus is isolated within a short interval of time [102].

Detection of “orphan polioviruses” or virus that are not closely linked to previously sequenced isolates indicates gaps or suboptimal surveillance. The length of the gap is roughly proportional to 1% single nucleotide divergence per year [39]. Orphan viruses [91, 105, 191] were responsible for most cVDPV outbreaks. This is in contrast to the situation in Nigeria where intensive AFP

surveillance triggered by the circulating wild polioviruses also revealed the initial stages in circulation of multiple lineages of predominantly type 2 cVDPVs [106]. The presence of type 2 cVDPVs, iVDPVs, and aVDPVs is of concern since despite elimination of transmission of wild type 2 in 1999, neurovirulent serotype 2 poliovirus is still among us [124, 128].

A number of countries that switched from OPV or combinations of OPV and IPV to exclusive use of IPV conducted environmental surveillance for OPV after the transition (reviewed in [143]). The OPV rapidly disappeared from the environmental samples but imported OPV-like isolates have been isolated from time to time presumably imported from OPV-using countries [233].

One of the important milestones toward achieving eradication is the containment of all potential sources of the pathogen. In 1999, a process for containing all laboratory stocks of wild poliovirus was initiated by the World Health Assembly entitled *Global action plan for the laboratory containment of wild polioviruses* or GAP I (WHO/V&B/99.32). A revised plan, GAP II, included two phases: (1) the identification in all facilities of all known poliovirus stocks and any material that could potentially contain live wild poliovirus, for example, any stool specimens that were collected at times when poliovirus was endemic, and (2) the containment of these stocks by destroying them, rendering them noninfectious, or transferring them to a minimal number of laboratories certified by the WHO as having appropriate BSL3/polio biosafety facilities and justification to work with wild viruses. A draft of the next version, GAP III, extends containment of wild polio to now also include containment of OPV/Sabin strains, and concentrates on minimizing risks associated from facilities that work with polioviruses and vaccine production and storage facilities after eradication of wild poliovirus transmission and cessation of OPV vaccination. Pathways of exposure from these facilities and assessment of the risks from a literature review have been calculated [134]. After risk analysis, a goal was set to reduce the number of such facilities globally to <20 essential facilities that meet required safeguards.

The original target date for polio eradication was not met. By 2001, the WHO Global Commission for the Certification of the Eradication of Poliomyelitis extended eradication to include elimination of

circulating VDPVs (Certification of the Global Eradication of Poliomyelitis Report of the sixth meeting of the Global Commission for the Certification of the Eradication of Poliomyelitis. Washington D.C., 28–29 March 2001 WHO/V&B/01.15). The current target date for interruption of all wild poliovirus has been moved to 2013 (Global Polio Eradication Initiative – Programme of Work 2009 and financial resource requirements 2009–2013. WHO/POLIO/09.02). This section will conclude with a discussion of the three major problems that have accounted for the delay in completing the GPEI, “failure to vaccinate,” “vaccine failure,” and the emergence of “vaccine-derived viruses.”

Among the reports available online at the WHO website for polio eradication, [www.polioeradication.org](http://www.polioeradication.org), is a report on the annual percentage of children in each country who received a minimum of three doses of polio vaccine annually since 1980. This report provides a complete picture of current polio immunization status. However the variability in coverage between countries and the annual fluctuation within countries illustrates the problem of failure to vaccinate and is also an indication of problems in sustaining the high coverage necessary for successful eradication.

Wild and vaccine-derived poliovirus can penetrate and circulate within areas where vaccination coverage is low or where vaccination has been discontinued [106, 117, 234]. When this occurs this is an example of “failure to vaccinate.” The temporary cessation of vaccination in Nigeria in 2003 [235, 236] is usually presented as the classic example for the consequences of a failure to vaccinate. The situation was complicated by suboptimal coverage within Nigeria when vaccination resumed within 12 months and the suboptimal coverage in other countries that had person-to-person contacts with infected Nigerians. Thirteen countries with 52% coverage have had multiple introductions of wild poliovirus from Nigeria, while another eight countries with higher 83% coverage did not have repeated outbreaks [1]. Use of type 1 mOPV was successful in reducing the number of cases from wild type 1 [190, 234] but the decreased use of tOPV led to a significant increase in cases due to wild type 3. Subsequent use of mOPV1, mOPV3, and bOPV has effectively reduced the number of cases due to wild types 1 and 3 [237]. Unfortunately suboptimal

immunization with any vaccine that contained type 2, presented fertile conditions for the emergence of multiple lineages of type 2 CVDPVs some of which have continued to circulate well into 2010 [106] (see discussion on page 8157, 8161).

A 2010 outbreak that started from a wild type 1 virus imported into Tajikistan from India spread into the Russian Federation. This was the first outbreak due to wild poliovirus in the WHO European region since it was declared polio-free in 2002 [221]. Again failure to vaccinate with coverage sufficient to maintain herd immunity was the main factor that facilitated the outbreak. As of June 2010, wild poliovirus cases from this outbreak accounted for more than 70% of all wild polio cases reported in 2010. Four NIDS with mOPV1 were conducted since the start of the outbreak.

When poliomyelitis occurs in vaccinated individuals it is categorized as “vaccine failure.” Current reports on India (20th Meeting of the India Expert Advisory Group for Polio Eradication Delhi, India, 24–25 June 2009 [www.polioeraication.org](http://www.polioeraication.org)) indicate that most of India is poliomyelitis-free with the exception of the north where both wild type 1 and type 3 still circulate. Type 1 and 3 mOPV have helped to reduce the number of circulating lineages and to constrict the areas within which the wild polioviruses are still circulating and causing cases. However, lack of cases in the south does not mean absence of wild poliovirus as environmental surveillance has revealed silent wild polio in Mumbai. The reservoir is not only a problem for India. Populations with suboptimal coverage in other countries are also at risk, as shown by the 2010 outbreak in Tajikistan and the Russian Federation that was traced back to northern India [221].

Vaccine failure in children in India refers to the finding that antibody response or seroconversion in children required more than the recommended three doses of tOPV [238, 239]. In Uttar Pradesh and Bihar in north India, local conditions exist where even administration of five doses of OPV does not induce the expected seroconversion rates. Approximately 15 doses were required to reach population immunity [240]. The fact that the age when the disease is acquired had not shifted upward was taken as an additional indication of vaccine failure [1]. Various trials of efficacy of mOPVs and bOPV and fractional IPV are

underway to evaluate their short-term effectiveness in halting endemic transmission and their long-term performance in maintaining protective titers. In India mucosal immunity in response to vaccination with OPV varied depending on location, serotype, and vaccine formulation [241].

The high number of additional doses needed to achieve herd immunity in some regions such as in northern India combined with the additional cost of OPV in annual and sub-annual vaccination campaigns must be taken into account when comparing the cost effectiveness of OPV and IPV in inducing effective herd immunity.

A less serious type of vaccine failure is based on observations that do not completely confirm the paradigm that OPV prevents replication during subsequent exposures to poliovirus. Israeli children who had concluded a primary immunization schedule consisting of three doses of OPV and three doses of IPV by 18 months of age had seroconverted for all three serotypes with geometric mean titers  $>1,000$  [159]. One month after the last vaccination they were challenged with an additional dose of OPV. Up to 60% of children excreted at least one OPV serotype between 1 and 3 weeks, the upper range of similar studies reviewed in that report. There was no evidence of transmission to siblings or mothers of these children, most likely because of good hygiene [159]. These rates were comparable to other similar studies [159]. Hygiene and high coverage probably also contributed to the fact that there was also no evidence for OPV circulation in IPV-vaccinated populations in the United States living adjacent to OPV-vaccinated populations in Mexico [242].

A number of comprehensive reviews on vaccine-derived polioviruses have been published [3, 62, 79, 105, 114, 243–246]. As described above (page 15), vaccine-derived polioviruses evolve either during person-to-person transmission (cVDPVs) or during relatively rare [247] persistent infections in immunodeficient patients (iVDPVs) (see reviews [3, 4, 105, 114]). To date (2010) there have been 12 cVDPV outbreaks [106]. Using the definition of outbreak in the context of eradication (i.e., even a single case), the number of outbreak may be even higher. For example, in 2009, 21 cases due to cVDPVs were found in four countries in addition to 153 in Nigeria and a case in Guinea traced back to Nigeria [248], and the cases in

Nigeria represent emergence of multiple independent lineages [106].

Most outbreaks caused by cVDPVs have been caused by a single lineage that spread rapidly through a susceptible cohort within the general population. The outbreaks were only discovered after silent circulation of the VDPVs for more than 1 year or more indicating gaps in surveillance. By the time such outbreaks became evident, the genomes of the isolates had usually recombined with those of the progeny of other vaccine serotypes or closely related non-polio enteroviruses. When OPV is introduced or reintroduced into a population with cohorts of naïve individuals as in Nigeria, adequate surveillance revealed that in early stages of reemergence more than one independent lineage may emerge [106]. There is also a potential for recombination of cVDPVs with wild-type viruses in areas where both co-circulate as shown from retrospective molecular analysis of isolates during endemic circulation of wild polioviruses [82]. Luckily from the point of view of eradication, cVDPV outbreaks resemble outbreaks of poliomyelitis from wild polioviruses introduced into polio-free areas and their chains of transmission can be broken by similar vaccination responses [106, 234].

In certain circumstances poliovirus can establish persistent infection in immunodeficient individuals. The types of immune deficiencies of known chronic excretors have been reviewed [3, 4, 62, 105]. The genomes of the Sabin strains are unstable [114] and reversion of nucleotide changes that attenuated neurovirulence appear even among the progeny virus excreted by primary OPV vaccinees. These reversions are believed to improve the replicative fitness of the isolates [123] and are responsible for the rarity of vaccine-associated paralytic poliomyelitis cases (VAPP; 1 per 500,000–1,000,000 vaccinations of naïve infants, and 7,000 times higher for some immunodeficient individuals) and cVDPV outbreaks [3, 105]. Thus it is not surprising that reversion of attenuation also occurs at an early stage in chronic excretors [62]. During 4 months of observation of long-term excretion in a healthy child [62] type 1 virus diverged by 1.1% and evolved toward full reversion to wild-type phenotypic properties similar to the Mahoney parent of the Sabin 1 strain. It is less obvious why these isolates so quickly predominate in the quasispecies of persistently infected

immunodeficient individuals. The process by which persistence is established and maintained may present selection through bottlenecks within a single individual that is similar to the bottleneck by which only a single progeny or a subpopulation of the quasispecies is passed onto the next host in person-to-person transmission among immune competent hosts. Selection by passage through bottlenecks was also suggested to explain evolution of wild poliovirus during long-term expression [249]. Examination of the genomes of iVDPV isolates differentiates them from cVDPVs in that significantly fewer heterotypic recombinations occur [4, 110] and intrageneous recombination appears to be largely absent [4, 105]. Interestingly, more than one highly divergent lineage may be recovered from a single stool sample from persistently infected individuals [3, 110]. This suggests that persistence and evolution occur in separate sites although intratypic recombination indicates that some mixed infections in a single cell must occur. Only a single serotype was detected in most (30 of 33) long-term excretors identified between 1962 and 2006 [4] and this pattern has continued to date. Amino acid changes in capsid proteins may allow polo to establish persistence in cells of the CNS [250].

Molecular analysis of phylogenetically related highly diverged (>10%) aVDPVs isolated from sewage in Finland, Slovakia, and Israel reveals a pattern of amino acid substitutions in or near neutralizing antigenic epitopes and lack of intrageneous recombination that resembles the pattern of changes in iVDPVs and is qualitatively different from evolutionary changes in cVDPVs [109, 251]. This pattern and the extended periods of time over which phylogenetically related polioviruses have been isolated from the same sewage systems and surveillance sites within those systems strongly suggests that replication of the related viruses has taken place in one immunodeficient host, or a very limited number of individuals in contact with such a host. Routine monthly sewage surveillance of catchment areas representing 35–40% of the population in Israel intermittently and repeatedly revealed the presence of highly diverged type 2 VDPVs 2 between 1998 and 2010. Phylogenetic analysis indicated that the isolates came from two different and unrelated persistent infections. Isolates from one foci have been isolated intermittently for 12 years and the second for 4 years.

In addition there was a single, respectively, and a single isolation of a highly diverged type 1 VDPV. The situation in Finland is particularly interesting and unusual, since evidence suggests that the infected individual is simultaneously and persistently infected with three highly diverged VDPV serotypes [251].

Most mutations in iVDPVs and aVDPVs are synonymous and are observed in third position codons. These synonymous substitutions occur at similar rates to those for poliovirus during person-to-person transmission [39, 101, 107, 109]. Non-synonymous amino acid substitutions affect antigenicity, neurovirulence, receptor binding motifs, hydrophobic pockets, and drug sensitivity.

The prevalence of aVDPV excretors is unknown, but additional countries with excretors of aVDPVs are being reported as environmental surveillance is introduced into more and more regions [143, 246, 252–256]. Hovi et al. [143] have proposed expanding the suggestion that the GPLN include regular monitoring for cVDPVs [257] through increasing the number of laboratories that employ routine environmental surveillance. It is important to determine the exact nature of the immune status of these types of persistent excretors since it may be different than that for identified persistently infected individuals. Unfortunately the individuals infected with these aVDPVs remain unidentified, and will most likely remain so for a long time [143]. The most frustrating attempt to locate such an excretor occurred in Slovakia where moving sampling sites progressively upstream successively restricted the excretor to a population of 500 individuals before detection ceased [143].

There is no consensus on the extent that persistent VDPVs may affect the realization of eradication [133, 163]. Determining the number of unidentified persistent infections is becoming more urgent as eradication of wild polio approaches (WHO 16th Informal Consultation Of The Global Polio Laboratory Network, 22–23 September 2010, Geneva, Switzerland, WHO/HQ. Summary of Discussions and Recommendations). Some researchers believe that VDPVs may pose an insurmountable problem [114, 127] while others feel that the problem is less severe [3, 4, 126]. Most of the identified persistent excretors had primary B-cell-related immunodeficiencies [4, 105]. While molecular epidemiological analysis has indicated that

highly diverged neurovirulent anonymous VDPVs isolated from sewage in Finland, Slovakia, and Israel [109, 233, 258] resemble the molecular epidemiology of poliovirus isolates excreted over time by identified excretors of iVDPVs, the exact nature of the immune status that has presumably enabled infection to persist in these aVDPV excretors remains unknown. The time course of excretions, the rate of nucleotide substitutions in virus isolated from identified persistent excretors, and genomic recombination patterns have been consistent with the establishment of persistence and evolution of the virus in these individuals rather than transmission of an iVDPV. Highly diverged iVDPVs (as opposed to cVDPVs and less diverged iVDPVs) have not been found during routine AFP surveillance of cases of immune competent individuals [3]. One clear indication that iVDPV-like aVDPVs are transmissible comes from a study of silent transmission in infected children in an under-vaccinated community in Minnesota [259] where 8 of 23 infants had evidence of type 1 poliovirus of VDPV infection. While this absence of documented transmissibility of very highly diverged VDPVs is encouraging, it may only be circumstantial, since most of the highly diverged neurovirulent aVDPVs have been found in the environment of countries with high vaccine coverage and good hygiene barriers that have also prevented circulation and appearance of wild poliovirus cases even after neurovirulent wild polio was introduced from external reservoirs [102].

The amino acid substitutions in neutralizing antigenic epitopes/receptor binding residues in iVDPVs and iVDPV-like aVDPVs may have helped specialize these virus isolates for microenvironments within the gut during persistent infections. These same changes might affect/reduce transmission via the oral–oral route in communities where there is high vaccine coverage and good hygiene. If true, this might significantly reduce the threat to eradication, despite the highly neurovirulent nature of these isolates in animal model systems and the decreased geometric mean neutralizing antibody titers against some of these excreted iVDPV and aVDPV isolates in the general public in the highly immunized communities where these isolates are found [109, 233]. It must also be taken into account that identified and unknown excretors are free to travel to communities with poor vaccine coverage and

substandard hygiene where the fecal–oral transmission route is more important and continuous person-to-person transmission easier to maintain.

### **Future Directions: The Endgame Stage of Eradication and Sustainability of Postpolio Eradication**

This section will start with an overview of current accomplishments to provide a suitable background for the discussion of future directions and sustainability. One of the best online sources for keeping up to date on eradication can be found at [www.polioeradication.org](http://www.polioeradication.org).

There has been a >99% overall reduction in the number of cases since adoption of the Global Poliovirus Eradication Initiative in 1988 [248]. The Region of the Americas (AMR) was certified to be free from all three serotypes of indigenous wild polioviruses in 1994 [260, 261] and the last case anywhere in the world from wild type 2 polio occurred in India in 1999 [186, 262]. Subsequently the Western Pacific Region (WPR) in 2000 [263] and the European Region (EUR) in 2002 [264] have also been certified to be free from indigenous wild polioviruses. These successes have been due to the dynamic nature of the eradication program where vaccination strategies have been adapted in response to specific problems and to changing conditions emerging as the endgame approached [1]. Four countries remain where indigenous poliovirus has continued, Afghanistan and Pakistan in the WHO Eastern Mediterranean Region (EMR), Nigeria in the WHO African Region (AFR), and India in the WHO South-East Asia Region (SEAR).

The accumulated costs for the vaccination program have exceeded 4.5 billion US dollars. National governments (list by alphabetical order: Australia, Austria, Belgium, Canada, Denmark, Finland, Germany, Ireland, Italy, Japan, Luxembourg, the Netherlands, Norway, the United Kingdom, and the United States) have provided a significant portion of the necessary funding. NGOs (WHO, UNICEF, Rotary International, the Bill and Melinda Gates Foundation, and the International Red Cross and Red Crescent societies), the World Bank, and corporate partners (Aventis Pasteur, De Beers) have also made significant contributions toward purchase of vaccines and for applied

and basic polio research. In addition to paid professional staff, more than ten million volunteers have assisted in the global vaccination program. Their knowledge of local practices and beliefs has provided a significant asset to the GPEI [1].

One of the goals of eradication was to reach a stage where vaccination could be discontinued, as was the case for smallpox vaccinations after eradication of smallpox [265]. The estimated annual savings would be enormous and could be used to fund other global health initiatives. Similarly the organizational capabilities experience expertise and facilities of member Laboratories in the Global Polio Laboratory Network would also be employed to solve other health-related problems. The three main problems that have delayed eradication originally scheduled for 2000 have been discussed. Among these problems, chronic excretion of vaccine-derived viruses probably remains the most serious obstacle since the number of excretors remains unknown and there are no universally recognized methods of curing chronic excretion in those chronic excretors who have been identified.

Between January 2009 and June 2010, the Global Polio Laboratory Network analyzed 258,000 fecal specimens from 130,000 AFP cases for the presence of poliovirus, and between January 2009 and September 2010 it detected introductions of wild poliovirus into 23 previously “polio-free countries,” countries where indigenous polio transmission had been interrupted. Nineteen were in the African region and included seven countries (Burkina Faso, Benin, Chad, Côte d’Ivoire, Mauritania, Niger, and Togo) where the outbreak isolates were related to previous importations into those countries as was the case for Sudan in the Eastern Mediterranean Region. One of the more serious setbacks for eradication was the introduction of wild poliovirus into Tajikistan from Uttar Pradesh in India marking the first outbreak in the European region since it was certified poliovirus-free in 2002 [221]. The large outbreak (>450 confirmed cases) ensued spread into the Russian Federation and resulted in an immediate tenfold increase in the amount of samples that needed to be processed by the Polio Laboratories in the region. In all of these importation countries and/or regions, large-scale coordinated SIAs were conducted. The spread of wild poliovirus to poliovirus-free countries from Nigeria and India via Tajikistan illustrate the need

to maintain high population immunity until all transmission of wild virus has ceased. Similarly, the emergence of multiple lineages of neurovirulent VDPVs in Nigeria and the increasing frequency of isolation of aVDPVs as more countries adopt environmental surveillance reinforce the need to discontinue use of OPV globally in a coordinated effort or staged manner. See Ehrenfeld et al. [266] for a review and discussion of key issues that have affected and will affect the GPEI, including: safety for volunteers in areas of strife, the low efficiency of OPV to induce herd immunity in certain settings, the requirement for maintaining high coverage even after eradication, the inherent mutability of OPV, problems for establishing the safety and efficacy and costs of new vaccines, new vaccine formulations, and scaling up alternatives to OPV.

The saying “*May you live in interesting times,*” often attributed to an ancient Chinese proverb or curse, appears appropriate for describing the current status in the quest to eradicate wild polioviruses. Eradication, which is tantalizingly close, will require substantial changes in vaccination policy and practice [117]. It must also include appropriate emergency response measures to control reemergence. “The ideal vaccine choice for the stockpile should be effective in any outbreak scenario, protect all vaccinees with one dose, spread to and protect the unvaccinated population, and have no detrimental effect” [267]. Long-term effects should be considered. While mOPV might be the most effective in rapidly controlling an outbreak and spreading and protecting unvaccinated individuals [267], plans that preferably do not require use of OPV adjacent to areas with high concentrations of unvaccinated individuals would be better in the long run [117, 268]. The reader is referred to the website of poliovirus eradication ([www.polioeradication.org](http://www.polioeradication.org)) for the latest information on past, current, and future policies.

Three problems have delayed the realization of eradication as has been discussed. Currently available vaccines can overcome “failure to vaccinate,” provided that enough doses of vaccines are made available, that there is the political will to use them, and that natural or man-made disasters do not prevent reaching the children for vaccination. Preliminary results from newly approved monovalent and bivalent oral polio vaccines and clinical studies using fractional doses of

IPV indicate that there may already be a solution for “vaccine failure” which is exemplified by the poor seroconversion rates for OPV in northern India [240]. The third major problem, “vaccine-derived polioviruses” is more complex, since VDPVs can evolve by person-to-person transmission (cVDPVs) or during persistent infections (iVDPVs). The spread of cVDPVs can be interrupted using the same methods as used to stop transmission of wild poliovirus (paradoxically including use of OPV in vaccination campaigns), since cVDPVs behave like wild virus [106, 234]. Moreover, while it is easy to say that current GPEI plans to coordinate global cessation of the use of OPV will prevent VAPP [206] and emergence of new cVDPVs, at this juncture the actual process is quite complicated and associated with a number of risks. The main problem that will need to be solved is the shedding of highly neurovirulent VDPVs into the environment for prolonged periods by identified and unidentified, persistently infected individuals. There is currently no universal solution to this problem [63, 125]. As long as shedding persists (perhaps as long as some of these individuals remain alive), containment as envisioned in GAP III will be incomplete and high vaccination coverage will need to be maintained. A related but more difficult problem that will need to be solved is (a) to determine the prevalence of unidentified, presumably persistently infected individuals who are responsible for shedding the highly diverged aVDPVs that have been isolated from environmental surveillance, and (b) to identify the presumably persistently infected individuals to determine the physiological conditions that enabled persistence and to try and clear the persistent infection with current or future antiviral treatments. As the endpoint of eradication of wild poliovirus is approached, the number of cases of poliomyelitis will decrease while the number of silent infections may increase as a result of high vaccine coverage. Under these conditions supplemental surveillance programs such as enterovirus surveillance and environmental surveillance will become an even more important tool for providing geographical information for designing NIDS, SNIDS, and final mopping-up campaigns for eradication and for monitoring for post-eradication reemergence.

“Although OPV has been the mainstay of the eradication program, its continued use is ironically

incompatible with the eradication of paralytic disease (since) vaccine-derived viruses consistently emerge as a consequence of the inherent genetic instability of poliovirus [122].” “Eradication of vaccine” suggested in 1997 [204, 205] has become recommended policy on condition that provisions of GAP III for safety in vaccine production and polio laboratories are met [264, 269]. A model describing the impact of cVDPVs on eradication indicated that the probability of eradicating polio with continuous use of OPV was not very likely [270]. Alternative vaccination should be continued during and especially after the transition to maintain high coverage and to avoid the buildup of large susceptible populations during the time when there is the highest risk for reemergence of OPV strains [117, 134, 270–272]. Low population immunity remains the main known risk factor for the emergence and spread of cVDPVs [234, 243]. Since most of the cVDPVs in outbreaks circulated silently for months or years (VP1 divergence >2%) before detection in AFP cases, it is imperative that surveillance be improved and expanded to high-risk regions to detect silent circulation of VDPVs as early as possible.

Endgame vaccination strategies have been reviewed [3, 122, 163, 266, 268] and include (1) indefinite use of OPV, (2) cessation of all polio immunization (3) transition to use of IPV, by synchronous coordinated cessation of all use of OPV with (a) limited use of IPV or (b) replacement of all OPV with IPV, (4) country-by-country cessation of OPV use with options (3a) or (3b), (5) sequential removal of Sabin strains from OPV, as eradication proceeds, (6) development of new vaccines, and (7) indefinite use of IPV or new vaccines. The synchronous cessation of OPV has several problems particularly if inexpensive alternatives are not in place when it occurs, since this vacuum may result in large populations of naïve individuals, in whom, polio could reemerge, after periods of silent circulation, with high force and rapid spread. Such a scenario also does not address the potential risks of unidentified chronic excretors. A gradual shift to IPV may avoid some of the programmatic disadvantages that coordinating a synchronous shift would have on vaccination programs and vaccination production facilities. It also potentially provides a longer window for industry to increase production, integrate information from current fractional vaccine dosage and alternative routes of

administration trials, and overcome problems of biocontainment and antigenicity associated with optional use of killed OPV as a substitute for the wild strains used in IPV production.

The WHO and UNICEF regularly consult informally with vaccine manufacturers to discuss the implications and practicality of vaccine policy decisions (summaries are available from the Internet using variations of a search for “WHO/UNICEF Informal Consultation with IPV and OPV Manufacturers”). For example, the 3rd WHO/UNICEF Informal Consultation with IPV and OPV Manufacturers (2003) included a discussion of post-eradication needs and biocontainment requirements and the 5th (2006) included updated information on progress of the GPEI and OPV cessation strategies.

The financial requirements for the transition period are complicated and have been set forth by the WHO (WHO Global Polio Eradication Initiative – Programme of Work 2009 and financial resource requirements 2009–2013 WHO/POLIO/09.02). The bottom line is that alternatives to OPV must be affordable [234]. Three recent reports deal in depth with the economics and practicality of universal replacement of OPV with IPV: (a) Global Post-eradication IPV Supply and Demand Assessment: Integrated Findings, March 2009, and (b) The supply landscape and economics of IPV-containing combination vaccines: Key findings, May 2010, both commissioned by the Bill & Melinda Gates Foundation and prepared by Oliver Wyman, Inc., and (c) Improving the affordability of inactivated poliovirus vaccines (IPV) for use in low- and middle-income countries – An economic analysis of strategies to reduce the cost of routine IPV immunization, April 20, 2010, prepared for PATH by Hickling, Jones, and Nundy. The second report [273] presents a thorough review of the current options and risks for new vaccines and vaccine formulations for achieving and maintaining eradication. New generations of inactivated polio vaccines may need to be developed for post-eradication use [266, 274] and they may have to be used indefinitely.

A number of decisions must be made now, some based on incomplete knowledge, because of the long lead time needed between planning facilities and final production of regulatory agency-approved products. For example, while fractional doses significantly reduce



costs, they are less effective than full doses and there is little data on kinetics of waning, while questions still exist concerning sufficient antigenicity of Sabin IPV. Additional complications involve testing and regulatory approval of new products or formulations (see discussion on regulations and standardization of IPV and IPV combination vaccines in Baca-Estrada and Griffiths [275] and the views of vaccine producers [276, 277]). The good news is that when “new” polio vaccine, type 1 mOPV, was needed, it was produced by two companies and licensed in three countries in a relatively short time, 6 months [135, 278, 279]. (Quotation marks were used around the word new since in actuality millions of monovalent doses of each serotype had been used before introduction of tOPV [280] when old licenses had been left to expire). Licensing was also aided by the fact that monovalent batches were produced and safety tested before being combined to produce tOPV and only qualified tOPV producers were approached to provide mOPVs [279]. Ironically if mOPVs are more effective than respective serotypes in tOPV because of increased titers and longer replication times, the increased number of nucleotide substitutions may increase the potential for cVDPV outbreaks by the serotype used [103] or conversely from the remaining serotypes (or serotype if bOPV is used). Supporting this is the emergence of significantly higher numbers of type 1 viruses with increased antigenic divergence from Sabin 1 after a birth dose of mOPV1 and a second exposure to Sabin 1 [111]. Most (71%) were isolated from stools from infants who did not seroconvert after the birth dose [111]. Rapid licensing of bOPV on January 10, 2010, followed release of efficacy results on June 2009 (issue 6 PolioPipeline, summer 2010). The bad news is that combination vaccines containing IPV cannot be frozen raising questions about long-term stability and appropriate reference standards [275].

There have been a number of attempts to rationally redesign the sequence of vaccine seed strains to make them more stable and safer to use in vaccine production facilities in the post-eradication period [3]. One drawback is that there is no empirical data on how these new viruses will behave in the field especially in relation to genome stability and the ability to recombine with heterotypic or intragenic enteroviruses. Changing codons to equivalent but rarely used

synonymous codons based on studies of codon use bias or increasing the frequency of CpG and UpA dinucleotides are methods to change the substitution rate [97, 281, 282]. Others modifications have led to polioviruses that can grow in nonhuman cell lines for production but have very low ability to infect human cells.

Widespread vaccination will continue at least during the 3-year period between the last case due to wild poliovirus and certification that wild poliovirus transmission has been interrupted globally. However, global vaccination should be continued for much longer since by one model [283], after 3 years there would only be a 95% certainty that all silent circulation had in fact ceased and the probability after 5 years ranged between 0.1% and 1%, while a more recent model has predicted a very high probability of reemergence within 10 years after eradication by VDPVs or accidental release of virus from vaccine production facilities, a polio laboratory, or bioterror [272]. Consequently vaccination will need to continue for at least 10 years after eradication. A special issue of the journal *Risk Analysis* (Volume 26, Issue 6, 2006) has been devoted to risks associated with polioviruses before, during, and after eradication of wild poliovirus. Finally vaccination with IPV may be continued indefinitely at least in countries where aVDPVs continue to be isolated from the environment with attendant risk from a polio vaccine production facility operating in a polio-free era (see discussion above on GAP III). To reiterate, current contingency plans for use of OPV in response to reemergence need to be revised based on the data on circulation of live vaccine strains after temporary and/or partial cessation of vaccination [117].

The final and one of the most important problems that must be successfully dealt with is to answer the question: “How can current achievements and eradication be sustained once the endgame has been concluded?” Ideally a major public health undertaking such as eradication requires a cost-benefit analysis, sufficient funds at the beginning, the means for achievement at hand, and the political and social will to carry the process through to the end. Delays and problems with fund-raising, especially when they occur during the endgame, may derail the entire effort [1]. Some problems with sustainability are associated with management and not scientific problems [1, 284].

However, new unanticipated scientific problems may appear which further delay polio eradication. After all, awareness of the potential problems from cVDPVs in communities with low vaccine coverage and chronic excretors of VDPV primarily appeared during the endgame of eradication when the global burden of cases had decreased by >99% and only after appropriate analytic tools to easily document and confirm VDPV had become widely available [3]. An example is the revelation of the 10-year circulation of cVDPVs in Egypt starting in 1983 [285] by retrospective phylogenetic examination of VP1 sequences.

The means to prevent disease and contain the spread of virus transmission when it emerges are already in hand. Safer and more cost-effective measures are in the pipeline that include schedule reduction and fractional doses, adjuvant use, optimizing of processing, Sabin or modified Sabin IPV, and noninfectious IPV [119, 273]. Sustainability for achieving eradication will depend on vaccine policy decisions made today, on the length of time it takes to eliminate all wild poliovirus transmissions, on political will and advocacy, on the motivation of volunteers and the level of local community involvement, program-related fatigue, and on the absence of complications [286] from bio-error, bioterror, or mother nature [291]. However the major determining factor will probably be the availability of financial resources [135]. Limited resources mean competition between routine immunization and eradication efforts during endgame. A predicted 1.3 billion USD funding gap in June 2010 is already forcing a reprioritization of planned activities (Global Polio Eradication Initiative Monthly Situation Report June 2010 [www.polioeradication.org](http://www.polioeradication.org).) “Even if there are no competing health needs, it is unlikely that immunization could be maintained indefinitely against a non-existent disease at a level that is sufficient to prevent vaccine-derived viruses evolving to cause epidemics” [163]. Programmatic setbacks such as those associated with failure to vaccinate (Nigeria and Tajikistan), vaccine failure (northern India), the frequency of repeated vaccination campaigns, or post-eradication reemergence must not be allowed to derail the current momentum and lead to program-related fatigue [1]. Detailed planning must be made for any post-eradication outbreaks (see Jenkins and Modlin [267] and Tebbins et al. [268]) and provisions to

implement them including stockpiling must be in place. Finally it is well worth reading “The pathogenesis of poliomyelitis: what we don’t know” by Nathanson [287] and “Gaps in scientific knowledge for the post-eradication world” by Minor [288].

## Note

Polio is the second human pathogen for which there is an ongoing global program for eradication that has reached the endgame. The first, smallpox, successfully completed the endgame and is now in the stage of post-eradication sustainability. Bioterror is the main threat to sustainability of smallpox eradication. This chapter will describe some of the difficulties with completing the endgame of polio eradication and then in sustaining postpolio eradication. More than the usual number of items are included in the glossary to make it easier for the reader to follow the progress of eradication as it unfolds in the large number of official documents that deal with a global eradication program and which contain the usual copious number of professional acronyms.

## Bibliography

1. Lahariya C (2007) Global eradication of polio: the case for “finishing the job”. *Bull World Health Organ* 85(6):487–492
2. Daniel TM, Robbins FC (1997) A history of poliomyelitis. In: Daniel TM, Robbins FC (eds) *Polio*. University of Rochester Press, Rochester, pp 5–22
3. Kew OM, Sutter RW, de Gourville EM, Dowdle WR, Pallansch MA (2005) Vaccine-derived polioviruses and the endgame strategy for global polio eradication. *Annu Rev Microbiol* 59:587–635
4. Sutter RW, Kew OM, Cochi SL (2008) Poliovirus vaccine – live. In: Plotkin SA, Orenstein WA, Offit PA (eds) *Vaccines*, 5th edn. W.B. Saunders/Elsevier, Philadelphia, pp 631–685
5. Burnet FM, MacNamara J (1931) Immunological differences between strains of poliomyelitic virus. *Br J Exp Pathol* 12:57–61
6. Wyatt HV (1985) Provocation of poliomyelitis by multiple injections. *Trans R Soc Trop Med Hyg* 79(3):355–358
7. Enders J, Weller T, Robbins FC (1949) Cultivation of the Lansing strain of poliomyelitis virus in cultures of various human embryonic tissues. *Science* 109(2822):85–87
8. Dulbecco R, Vogt M (1954) Plaque formation and isolation of pure lines with poliomyelitis viruses. *J Exp Med* 99(2):167–182
9. van Wezel AL (1967) Growth of cell-strains and primary cells on micro-carriers in homogeneous culture. *Nature* 216(5110):64–65
10. Cockburn WC (1988) The work of the WHO consultative group on poliomyelitis vaccines. *Bull World Health Organ* 66(2):143–154

11. Mendelsohn CL, Wimmer E, Racaniello VR (1989) Cellular receptor for poliovirus: molecular cloning, nucleotide sequence, and expression of a new member of the immunoglobulin superfamily. *Cell* 56(5):855–865
12. Dragunsky E, Nomura T, Karpinski K, Furesz J, Wood DJ, Pervikov Y et al (2003) Transgenic mice as an alternative to monkeys for neurovirulence testing of live oral poliovirus vaccine: validation by a WHO collaborative study. *Bull World Health Organ* 81(4):251–260
13. Wood DJ, Hull B (1999) L20B cells simplify culture of polioviruses from clinical samples. *J Med Virol* 58(2):188–192
14. Brodie M (1935) Active immunization against poliomyelitis. *Am J Public Health Nations Health* 25(1):54–67
15. Kolmer JA (1936) Vaccination against acute anterior poliomyelitis. *Am J Public Health Nations Health* 26(2):126–135
16. Salk JE (1953) Studies in human subjects on active immunization against poliomyelitis. I. A preliminary report of experiments in progress. *J Am Med Assoc* 151(13):1081–1098
17. Salk JE, Bazeley PL, Bennett BL, Krech U, Lewis LJ, Ward EN et al (1954) Studies in human subjects on active immunization against poliomyelitis. II. A practical means for inducing and maintaining antibody formation. *Am J Public Health Nations Health* 44(8):994–1009
18. Koprowski H, Jervis G, Norton T (1952) Immune responses in human volunteers upon oral administration of a rodent-adapted strain of poliomyelitis virus. *Am J Epidemiol* 55(1):108–126
19. Koprowski H (2006) First decade (1950–1960) of studies and trials with the polio vaccine. *Biologicals* 34(2):81–86
20. Sabin AB (1957) Properties and behavior of orally administered attenuated poliovirus vaccine. *J Am Med Assoc* 164(11):1216–1223
21. Cox HR, Cabasso VJ, Markham FS, Moses MJ, Moyer AW, Roca-Garcia M et al (1959) Immunological response to trivalent oral poliomyelitis vaccine. *Br Med J* 2(5152):591–597
22. Furesz J (2006) Developments in the production and quality control of poliovirus vaccines – historical perspectives. *Biologicals* 34(2):87–90
23. John TJ (2001) Anomalous observations on IPV and OPV vaccination. In: Brown F (ed) *Progress in polio eradication: vaccine strategies for the end game*. Karger, Basel, pp 197–208
24. Davidson I, Shulman LM (2008) Unraveling the puzzle of human anellovirus infections by comparison with avian infections with the chicken anemia virus. *Virus Res* 137(1):1–15
25. Racaniello VR et al (2007) Picornaviridae: the viruses and their replication. In: Chief E-I, Knipe DM, Howley PM, Editors A, Griffin DE, Lamb RA (eds) *Fields virology*, 5th edn. Wolters Kluwer/Lippincott Williams & Wilkins, Philadelphia, pp 795–838
26. Wimmer E, Hellen CU, Cao X (1993) Genetics of poliovirus. *Annu Rev Genet* 27:353–436
27. Wimmer E (2006) The test-tube synthesis of a chemical called poliovirus. The simple synthesis of a virus has far-reaching societal implications. *EMBO Rep* 7:53–59
28. Molla A, Paul AV, Wimmer E (1991) Cell-free, de novo synthesis of poliovirus. *Science* 254(5038):1647–1651
29. Poyry T, Kinnunen L, Hovi T (1992) Genetic variation in vivo and proposed functional domains of the 5' noncoding region of poliovirus RNA. *J Virol* 66(9):5313–5319
30. Pilipenko EV, Bliinov VM, Romanova LI, Sinyakov AN, Maslova SV, Agol VI (1989) Conserved structural domains in the 5'-untranslated region of picornaviral genomes: an analysis of the segment controlling translation and neurovirulence. *Virology* 168(2):201–209
31. Skinner MA, Racaniello VR, Dunn G, Cooper J, Minor PD, Almond JW (1989) New model for the secondary structure of the 5' non-coding RNA of poliovirus is supported by biochemical and genetic data that also show that RNA secondary structure is important in neurovirulence. *J Mol Biol* 207(2):379–392
32. Pelletier J, Sonenberg N (1988) Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature* 334(6180):320–325
33. Jang SK, Krausslich HG, Nicklin MJ, Duke GM, Palmenberg AC, Wimmer E (1988) A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. *J Virol* 62(8):2636–2643
34. Gamarnik AV, Andino R (1997) Two functional complexes formed by KH domain containing proteins with the 5' non-coding region of poliovirus RNA. *RNA* 3(8):882–892
35. Liu HM, Zheng DP, Zhang LB, Oberste MS, Kew OM, Pallansch MA (2003) Serial recombination during circulation of type 1 wild-vaccine recombinant polioviruses in China. *J Virol* 77(20):10994–11005
36. Racaniello VR (2001) Picornaviridae: the viruses and their replication. In: Fields BN, Knipe N, Howley P (eds) *Virology*, 4th edn. Lippincott Williams & Wilkins, Philadelphia, pp 685–722
37. Krausslich HG, Nicklin MJ, Lee CK, Wimmer E (1988) Polyprotein processing in picornavirus replication. *Biochimie* 70(1):119–130
38. Shulman LM, Handsher R, Yang CF, Yang SJ, Manor J, Vonsover A et al (2000) Resolution of the pathways of poliovirus type 1 transmission during an outbreak. *J Clin Microbiol* 38(3):945–952
39. Jorba J, Campagnoli R, De L, Kew O (2008) Calibration of multiple poliovirus molecular clocks covering an extended evolutionary range. *J Virol* 82(9):4429–4440
40. Bouchard MJ, Lam DH, Racaniello VR (1995) Determinants of attenuation and temperature sensitivity in the type 1 poliovirus Sabin vaccine. *J Virol* 69(8):4972–4978
41. Bergamini G, Preiss T, Hentze MW (2000) Picornavirus IRESes and the poly(A) tail jointly promote cap-independent translation in a mammalian cell-free system. *RNA* 6(12):1781–1790
42. Crawford NM, Baltimore D (1983) Genome-linked protein VPg of poliovirus is present as free VPg and VPg-pUpU in poliovirus-infected cells. *Proc Natl Acad Sci USA* 80(24):7452–7455
43. Kuhn RJ, Tada H, Ypma-Wong MF, Dunn JJ, Semler BL, Wimmer E (1988) Construction of a "mutagenesis cartridge" for poliovirus genome-linked viral protein: isolation and characterization of viable and nonviable mutants. *Proc Natl Acad Sci USA* 85(2):519–523

44. Ambros V, Pettersson RF, Baltimore D (1978) An enzymatic activity in uninfected cells that cleaves the linkage between poliovirus RNA and the 5' terminal protein. *Cell* 15(4):1439–1446
45. Wien MW, Chow M, Hogle JM (1996) Poliovirus: new insights from an old paradigm. *Structure* 4(7):763–767
46. Li Q, Yafal AG, Lee YM, Hogle J, Chow M (1994) Poliovirus neutralization by antibodies to internal epitopes of VP4 and VP1 results from reversible exposure of these sequences at physiological temperature. *J Virol* 68(6):3965–3970
47. Harkonen T, Lankinen H, Davydova B, Hovi T, Roivainen M (2002) Enterovirus infection can induce immune responses that cross-react with beta-cell autoantigen tyrosine phosphatase IA-2/IAR. *J Med Virol* 66(3):340–350
48. Page GS, Mosser AG, Hogle JM, Filman DJ, Rueckert RR, Chow M (1988) Three-dimensional structure of poliovirus serotype 1 neutralizing determinants. *J Virol* 62(5):1781–1794
49. Bodian D, Morgan IM, Howe HA (1949) Differentiation of types of poliomyelitis viruses; the grouping of 14 strains into three basic immunological types. *Am J Hyg* 49(2):234–245
50. Harber J, Bernhardt G, Lu HH, Sgro JY, Wimmer E (1995) Canyon rim residues, including antigenic determinants, modulate serotype-specific binding of polioviruses to mutants of the poliovirus receptor. *Virology* 214(2):559–570
51. Hogle JM, Chow M, Filman DJ (1985) Three-dimensional structure of poliovirus at 2.9 Å resolution. *Science* 229(4720):1358–1365
52. Filman DJ, Syed R, Chow M, Macadam AJ, Minor PD, Hogle JM (1989) Structural factors that control conformational transitions and serotype specificity in type 3 poliovirus. *EMBO J* 8(5):1567–1579
53. Koike S, Ise I, Nomoto A (1991) Functional domains of the poliovirus receptor. *Proc Natl Acad Sci USA* 88(10):4104–4108
54. Belnap DM, McDermott BM Jr, Filman DJ, Cheng N, Trus BL, Zuccola HJ et al (2000) Three-dimensional structure of poliovirus receptor bound to poliovirus. *Proc Natl Acad Sci USA* 97(1):73–78
55. WHO t (1998) Scheme adopted for use for L20B cells. *Polio LaB network quarterly update* IV(4):1–2
56. Zurbriggen S, Tobler K, Abril C, Diedrich S, Ackermann M, Pallansch MA et al (2008) Isolation of sabin-like polioviruses from wastewater in a country using inactivated polio vaccine. *Appl Environ Microbiol* 74(18):5608–5614
57. Abe S, Ota Y, Koike S, Kurata T, Horie H, Nomura T et al (1995) Neurovirulence test for oral live poliovaccines using poliovirus-sensitive transgenic mice. *Virology* 206(2):1075–1083
58. Horie H, Koike S, Kurata T, Sato-Yoshida Y, Ise I, Ota Y et al (1994) Transgenic mice carrying the human poliovirus receptor: new animal models for study of poliovirus neurovirulence. *J Virol* 68(2):681–688
59. Ren RB, Costantini F, Gorgacz EJ, Lee JJ, Racaniello VR (1990) Transgenic mice expressing a human poliovirus receptor: a new model for poliomyelitis. *Cell* 63(2):353–362
60. Smyth M, Pettitt T, Symonds A, Martin J (2003) Identification of the pocket factors in a picornavirus. *Arch Virol* 148(6):1225–1233
61. Salvati AL, De Dominicis A, Tait S, Canitano A, Lahm A, Fiore L (2004) Mechanism of action at the molecular level of the antiviral drug 3(2H)-isoflavene against type 2 poliovirus. *Antimicrob Agents Chemother* 48(6):2233–2243
62. Martin J (2006) Vaccine-derived poliovirus from long term excretors and the end game of polio eradication. *Biologicals* 34(2):117–122
63. MacLennan C, Dunn G, Huissoon AP, Kumararatne DS, Martin J, O'Leary P et al (2004) Failure to clear persistent vaccine-derived neurovirulent poliovirus infection in an immunodeficient man. *Lancet* 363(9420):1509–1513
64. Domingo E, Martinez-Salas E, Sobrino F, de la Torre JC, Portela A, Ortin J et al (1985) The quasispecies (extremely heterogeneous) nature of viral RNA genome populations: biological relevance—a review. *Gene* 40(1):1–8
65. Kinnunen L, Huovilainen A, Poyry T, Hovi T (1990) Rapid molecular evolution of wild type 3 poliovirus during infection in individual hosts. *J Gen Virol* 71(Pt 2):317–324
66. Crotty S, Cameron CE, Andino R (2001) RNA virus error catastrophe: direct molecular test by using ribavirin. *Proc Natl Acad Sci USA* 98(12):6895–6900
67. Eigen M (2002) Error catastrophe and antiviral strategy. *Proc Natl Acad Sci USA* 99(21):13374–13376
68. Racaniello VR, Baltimore D (1981) Cloned poliovirus complementary DNA is infectious in mammalian cells. *Science* 214(4523):916–919
69. Pollard SR, Dunn G, Cammack N, Minor PD, Almond JW (1989) Nucleotide sequence of a neurovirulent variant of the type 2 oral poliovirus vaccine. *J Virol* 63(11):4949–4951
70. van der Werf S, Bradley J, Wimmer E, Studier FW, Dunn JJ (1986) Synthesis of infectious poliovirus RNA by purified T7 RNA polymerase. *Proc Natl Acad Sci USA* 83(8):2330–2334
71. Cello J, Paul AV, Wimmer E (2002) Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science* 297(5583):1016–1018
72. Melnick JL (1996) Current status of poliovirus infections. *Clin Microbiol Rev* 9(3):293–300
73. Gustin KE, Sarnow P (2002) Inhibition of nuclear import and alteration of nuclear pore complex composition by rhinovirus. *J Virol* 76(17):8787–8796
74. Fox MP, Otto MJ, McKinlay MA (1986) Prevention of rhinovirus and poliovirus uncoating by WIN 51711, a new antiviral drug. *Antimicrob Agents Chemother* 30(1):110–116
75. Zeichhardt H, Otto MJ, McKinlay MA, Willingmann P, Habermehl KO (1987) Inhibition of poliovirus uncoating by disoxaril (WIN 51711). *Virology* 160(1):281–285
76. Jurgens CK, Barton DJ, Sharma N, Morasco BJ, Ogram SA, Flanagan JB (2006) 2Apro is a multifunctional protein that regulates the stability, translation and replication of poliovirus RNA. *Virology* 345(2):346–357
77. Takegami T, Kuhn RJ, Anderson CW, Wimmer E (1983) Membrane-dependent uridylylation of the genome-linked protein VPg of poliovirus. *Proc Natl Acad Sci USA* 80(24):7447–7451
78. Eigen M (1993) Viral quasispecies. *Sci Am* 269(1):42–49

79. Gavrilin GV, Cherkasova EA, Lipskaya GY, Kew OM, Agol VI (2000) Evolution of circulating wild poliovirus and of vaccine-derived poliovirus in an immunodeficient patient: a unifying model. *J Virol* 74(16):7381–7390
80. Cherkasova E, Laassri M, Chizhikov V, Korotkova E, Dragunsky E, Agol VI et al (2003) Microarray analysis of evolution of RNA viruses: evidence of circulation of virulent highly divergent vaccine-derived polioviruses. *Proc Natl Acad Sci USA* 100(16):9398–9403
81. Guillot S, Caro V, Cuervo N, Korotkova E, Combiescu M, Persu A et al (2000) Natural genetic exchanges between vaccine and wild poliovirus strains in humans. *J Virol* 74(18):8434–8443
82. Dahourou G, Guillot S, Le Gall O, Crainic R (2002) Genetic recombination in wild-type poliovirus. *J Gen Virol* 83(Pt 12): 3103–3110
83. Zhang Y, Wang H, Zhu S, Li Y, Song L, Liu Y et al (2010) Characterization of a rare natural intertypic type 2/type 3 pentarecombinant vaccine-derived poliovirus isolated from a child with acute flaccid paralysis. *J Gen Virol* 91(Pt 2):421–429
84. Tao Z, Wang H, Xu A, Zhang Y, Song L, Zhu S et al (2010) Isolation of a recombinant type 3/type 2 poliovirus with a chimeric capsid VP1 from sewage in Shandong, China. *Virus Res* 150(1–2):56–60
85. Blomqvist S, Savolainen-Kopra C, Paananen A, El Bassioni L, El Maamoun Nasr EM, Firstova L et al (2010) Recurrent isolation of poliovirus 3 strains with chimeric capsid protein Vp1 suggests a recombination hot-spot site in Vp1. *Virus Res* 151(2): 246–251
86. Brown B, Oberste MS, Maher K, Pallansch MA (2003) Complete genomic sequencing shows that polioviruses and members of human enterovirus species C are closely related in the noncapsid coding region. *J Virol* 77(16):8973–8984
87. Jiang P, Faase JA, Toyoda H, Paul A, Wimmer E, Goralbalena AE (2007) Evidence for emergence of diverse polioviruses from C-cluster coxsackie A viruses and implications for global poliovirus eradication. *Proc Natl Acad Sci USA* 104(22): 9457–9462
88. Riquet FB, Blanchard C, Jegouic S, Balanant J, Guillot S, Vibet MA et al (2008) Impact of exogenous sequences on the characteristics of an epidemic type 2 recombinant vaccine-derived poliovirus. *J Virol* 82(17):8927–8932
89. Kirkegaard K, Baltimore D (1986) The mechanism of RNA recombination in poliovirus. *Cell* 47(3):433–443
90. Kew O, Morris-Glasgow V, Landaverde M, Burns C, Shaw J, Garib Z et al (2002) Outbreak of poliomyelitis in Hispaniola associated with circulating type 1 vaccine-derived poliovirus. *Science* 296(5566):356–359
91. Estivariz CF, Watkins MA, Handoko D, Rusipah R, Deshpande J, Rana BJ et al (2008) A large vaccine-derived poliovirus outbreak on Madura Island–Indonesia, 2005. *J Infect Dis* 197(3):347–354
92. Cherkasova EA, Korotkova EA, Yakovenko ML, Ivanova OE, Eremeeva TP, Chumakov KM et al (2002) Long-term circulation of vaccine-derived poliovirus that causes paralytic disease. *J Virol* 76(13):6791–6799
93. Chumakov KM, Norwood LP, Parker ML, Dragunsky EM, Ran YX, Levenbook IS (1992) RNA sequence variants in live poliovirus vaccine and their relation to neurovirulence. *J Virol* 66(2):966–970
94. Reyes GR (2001) Ribavirin: recent insights into antiviral mechanisms of action. *Curr Opin Drug Discov Devel* 4(5): 651–656
95. Georgescu MM, Balanant J, Ozden S, Crainic R (1997) Random selection: a model for poliovirus infection of the central nervous system. *J Gen Virol* 78(Pt 8):1819–1828
96. Pfeiffer JK, Kirkegaard K (2005) Increased fidelity reduces poliovirus fitness and virulence under selective pressure in mice. *PLoS Pathog* 1(2):e11
97. Burns CC, Campagnoli R, Shaw J, Vincent A, Jorba J, Kew O (2009) Genetic inactivation of poliovirus infectivity by increasing the frequencies of CpG and UpA dinucleotides within and across synonymous capsid region codons. *J Virol* 83(19):9957–9969
98. Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R (2006) Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* 439(7074):344–348
99. Tao B, Fultz PN (1995) Molecular and biological analyses of quasispecies during evolution of a virulent simian immunodeficiency virus, SIVsmmPBj14. *J Virol* 69(4):2031–2037
100. de la Torre JC, Holland JJ (1990) RNA virus quasispecies populations can suppress vastly superior mutant progeny. *J Virol* 64(12):6278–6281
101. Kew OM, Mulders MN, Lipskaya GY, de Silva E, Pallansch MA (1995) Molecular epidemiology of polioviruses. *Sem Virol* 6:401–405
102. Manor Y, Blomqvist S, Sofer D, Alfandari J, Halmut T, Abramovitz B et al (2007) Advanced environmental surveillance and molecular analyses indicate separate importations rather than endemic circulation of wild type 1 poliovirus in Gaza district in 2002. *Appl Environ Microbiol* 73(18):5954–5958
103. Boot HJ, Sonsma J, van Nunen F, Abbink F, Kimman TG, Buisman AM (2007) Determinants of monovalent oral polio vaccine mutagenesis in vaccinated elderly people. *Vaccine* 25(24):4706–4714
104. Kew OM, Sutter RW, Nottay BK, McDonough MJ, Prevots DR, Quick L et al (1998) Prolonged replication of a type 1 vaccine-derived poliovirus in an immunodeficient patient. *J Clin Microbiol* 36(10):2893–2899
105. Kew OM, Wright PF, Agol VI, Delpeyroux F, Shimizu H, Nathanson N et al (2004) Circulating vaccine-derived polioviruses: current state of knowledge. *Bull World Health Organ* 82(1):16–23
106. Jenkins HE, Aylward RB, Gasasira A, Donnelly CA, Mwanza M, Corander J et al (2010) Implications of a circulating vaccine-derived poliovirus in Nigeria. *N Engl J Med* 362(25):2360–2369
107. Martin J, Dunn G, Hull R, Patel V, Minor PD (2000) Evolution of the Sabin strain of type 3 poliovirus in an immunodeficient patient during the entire 637-day period of virus excretion. *J Virol* 74(7):3001–3010

108. Martin J, Odoom K, Tuite G, Dunn G, Hopewell N, Cooper G et al (2004) Long-term excretion of vaccine-derived poliovirus by a healthy child. *J Virol* 78(24):13839–13847
109. Shulman LM, Manor Y, Sofer D, Handsheer R, Swartz T, Delpeyroux F et al (2006) Neurovirulent vaccine-derived polioviruses in sewage from highly immune populations. *PLoS One* 1:e69
110. Yang CF, Chen HY, Jorba J, Sun HC, Yang SJ, Lee HC et al (2005) Intratypic recombination among lineages of type 1 vaccine-derived poliovirus emerging during chronic infection of an immunodeficient patient. *J Virol* 79(20):12623–12634
111. van der Sanden S, Pallansch MA, van de Kassteelle J, El-Sayed N, Sutter RW, Koopmans M et al (2009) Shedding of vaccine viruses with increased antigenic and genetic divergence after vaccination of newborns with monovalent type 1 oral poliovirus vaccine. *J Virol* 83(17):8693–8704
112. Yakovenko ML, Cherkasova EA, Rezapkin GV, Ivanova OE, Ivanov AP, Eremeeva TP et al (2006) Antigenic evolution of vaccine-derived polioviruses: changes in individual epitopes and relative stability of the overall immunological properties. *J Virol* 80(6):2641–2653
113. Domingo E, Diez J, Martinez MA, Hernandez J, Holguin A, Borrego B et al (1993) New observations on antigenic diversification of RNA viruses. Antigenic variation is not dependent on immune selection. *J Gen Virol* 74(Pt 10):2039–2045
114. Agol VI (2006) Vaccine-derived polioviruses. *Biologicals* 34(2):103–108
115. Blomqvist S, Savolainen C, Laine P, Hirttio P, Lamminsalo E, Penttila E et al (2004) Characterization of a highly evolved vaccine-derived poliovirus type 3 isolated from sewage in Estonia. *J Virol* 78(9):4876–4883
116. Shulman LM, Manor Y, Sofer D (2010) Poliovirus vaccine and vaccine-derived polioviruses. *N Engl J Med* 363(19):1870
117. Korotkova EA, Park R, Cherkasova EA, Lipskaya GY, Chumakov KM, Feldman EV et al (2003) Retrospective analysis of a local cessation of vaccination against poliomyelitis: a possible scenario for the future. *J Virol* 77(23):12460–12465
118. Gumedé N, Venter M, Lentsoane O, Muyembe-Tamfum J, Yogoletlo R, Puren A et al (2010) Identification of vaccine-derived polioviruses (VDPVs) in the DRC from 2005 to 2010. *Commun Dis Sur Bull* 8(3):43–45
119. Resik S, Tejada A, Lago PM, Diaz M, Carmenates A, Sarmiento L et al (2010) Randomized controlled clinical trial of fractional doses of inactivated poliovirus vaccine administered intradermally by needle-free device in Cuba. *J Infect Dis* 201(9):1344–1352
120. Mohammed AJ, AlAwaidy S, Bawikar S, Kurup PJ, Elamir E, Shaban MM et al (2010) Fractional doses of inactivated poliovirus vaccine in Oman. *N Engl J Med* 362(25):2351–2359
121. Nirmal S, Cherian T, Samuel BU, Rajasingh J, Raghupathy P, John TJ (1998) Immune response of infants to fractional doses of intradermally administered inactivated poliovirus vaccine. *Vaccine* 16(9–10):928–931
122. Ehrenfeld E, Glass RI, Agol VI, Chumakov K, Dowdle W, John TJ et al (2008) Immunisation against poliomyelitis: moving forward. *Lancet* 371(9621):1385–1387
123. Yakovenko ML, Korotkova EA, Ivanova OE, Eremeeva TP, Samoilovich E, Uhova I et al (2009) Evolution of the Sabin vaccine into pathogenic derivatives without appreciable changes in antigenic properties: need for improvement of current poliovirus surveillance. *J Virol* 83(7):3402–3406
124. Modlin JF (2010) The bumpy road to polio eradication. *N Engl J Med* 362(25):2346–2349
125. Buttinelli G, Donati V, Fiore S, Marturano J, Plebani A, Balestri P et al (2003) Nucleotide variation in Sabin type 2 poliovirus from an immunodeficient patient with poliomyelitis. *J Gen Virol* 84(Pt 5):1215–1221
126. Arita I, Nakane M, Fenner F (2006) Public health. Is polio eradication realistic? *Science* 312(5775):852–854
127. Chumakov K, Ehrenfeld E, Wimmer E, Agol VI (2007) Vaccination against polio should not be stopped. *Nat Rev Microbiol* 5(12):952–958
128. Shulman LM, Manor Y, Sofer D, Mendelson E (2009) Type 2 polio still in our midst. *Science* 324(5925):334
129. Furione M, Guillot S, Otelea D, Balanant J, Candrea A, Crainic R (1993) Polioviruses with natural recombinant genomes isolated from vaccine-associated paralytic poliomyelitis. *Virology* 196(1):199–208
130. Lipskaya GY, Muzychenko AR, Kutitova OK, Maslova SV, Equestre M, Drozdov SG et al (1991) Frequent isolation of intertypic poliovirus recombinants with serotype 2 specificity from vaccine-associated polio cases. *J Med Virol* 35(4):290–296
131. Georgescu MM, Delpeyroux F, Tardy-Panit M, Balanant J, Combiescu AA et al (1994) High diversity of poliovirus strains isolated from the central nervous system from patients with vaccine-associated paralytic poliomyelitis. *J Virol* 68(12):8089–8101
132. Minor PD, John A, Ferguson M, Icenogle JP (1986) Antigenic and molecular evolution of the vaccine strain of type 3 poliovirus during the period of excretion by a primary vaccinee. *J Gen Virol* 67(Pt 4):693–706
133. Shulman LM, Manor Y, Sofer D, Swartz T, Mendelson E (2006) Oral poliovaccine: will it help eradicate polio or cause the next epidemic? *Isr Med Assoc J* 8(5):312–315
134. Dowdle W, van der Avoort H, de Gourville E, Delpeyroux F, Desphande J, Hovi T et al (2006) Containment of polioviruses after eradication and OPV cessation: characterizing risks to improve management. *Risk Anal* 26(6):1449–1469
135. Wood DJ (2006) Polio vaccine: the first 50 years and beyond. Summary of the meeting and next steps. *Biologicals* 34(2):171–174
136. Horstmann DM, Mc CR, Mascola AD (1954) Viremia in human poliomyelitis. *J Exp Med* 99(4):355–369
137. Steigman AJ, Sabin AB (1949) Antibody response of patients with poliomyelitis to virus recovered from their own alimentary tract. *J Exp Med* 90(4):349–372

138. Bodian D, Paffenbarger RS Jr (1954) Poliomyelitis infection in households; frequency of viremia and specific antibody response. *Am J Hyg* 60(1):83–98
139. Racaniello VR, Ren R (1996) Poliovirus biology and pathogenesis. *Curr Top Microbiol Immunol* 206:305–325
140. Alexander JP Jr, Gary HE Jr, Pallansch MA (1997) Duration of poliovirus excretion and its implications for acute flaccid paralysis surveillance: a review of the literature. *J Infect Dis* 175(Suppl 1):S176–S182
141. Minor PD (1992) The molecular biology of poliovaccines. *J Gen Virol* 73(Pt 12):3065–3077
142. Mas Lago P, Caceres VM, Galindo MA, Gary HE Jr, Valcarcel M, Barrios J et al (2001) Persistence of vaccine-derived poliovirus following a mass vaccination campaign in Cuba: implications for stopping polio vaccination after global eradication. *Int J Epidemiol* 30(5):1029–1034
143. Hovi T, Shulman LM, van der Avoort H, Deshpande J, Roivainen M, de Gourville EM (2011) Role of environmental poliovirus surveillance in global polio eradication and beyond, a review. *Epidemiol Infect* 18:1–13
144. Cohen-Abbo A, Culley BS, Reed GW, Sannella EC, Mace RL, Robertson SE et al (1995) Seroreponse to trivalent oral poliovirus vaccine as a function of dosage interval. *Pediatr Infect Dis J* 14(2):100–106
145. Gromeier M, Wimmer E (1998) Mechanism of injury-provoked poliomyelitis. *J Virol* 72(6):5056–5060
146. Ramlow J, Alexander M, LaPorte R, Kaufmann C, Kuller L (1992) Epidemiology of the post-polio syndrome. *Am J Epidemiol* 136(7):769–786
147. Gonzalez H, Olsson T, Borg K (2010) Management of postpolio syndrome. *Lancet Neurol* 9(6):634–642
148. Ogra PL (1995) Comparative evaluation of immunization with live attenuated and inactivated poliovirus vaccines. *Ann N Y Acad Sci* 754:97–107
149. Ogra PL, Karzon DT (1969) Distribution of poliovirus antibody in serum, nasopharynx and alimentary tract following segmental immunization of lower alimentary tract with poliovaccine. *J Immunol* 102(6):1423–1430
150. Ogra PL, Karzon DT (1969) Poliovirus antibody response in serum and nasal secretions following intranasal inoculation with inactivated poliovaccine. *J Immunol* 102(1):15–23
151. Sofer D, Handscher R, Abramovitz B, Shilon K, Manor Y, Halmut T, et al (2008) Determining vaccination efficacy: is the current minimum anti-polio neutralization antibody titer of >1:8 against Sabin strains high enough? Meeting of the three division of the international union of microbiological societies 5–15 Aug 2008, Istanbul
152. Valtanen S, Roivainen M, Piirainen L, Stenvik M, Hovi T (2000) Poliovirus-specific intestinal antibody responses coincide with decline of poliovirus excretion. *J Infect Dis* 182(1):1–5
153. Nishio O, Sumi J, Sakae K, Ishihara Y, Isomura S, Inouye S (1990) Fecal IgA antibody responses after oral poliovirus vaccination in infants and elder children. *Microbiol Immunol* 34(8):683–689
154. Ogra PL, Fishaut M, Gallagher MR (1980) Viral vaccination via the mucosal routes. *Rev Infect Dis* 2(3):352–369
155. Samoilovich E, Roivainen M, Titov LP, Hovi T (2003) Serotype-specific mucosal immune response and subsequent poliovirus replication in vaccinated children. *J Med Virol* 71(2):274–280
156. Onorato IM, Modlin JF, McBean AM, Thoms ML, Losonsky GA, Bernier RH (1991) Mucosal immunity induced by enhance-potency inactivated and oral polio vaccines. *J Infect Dis* 163(1):1–6
157. Vidor E, Caudrelier P, Plotkin S (1994) The place of DTP/eIPV vaccine in routine paediatric vaccination. *Rev Med Virol* 4(4):261–277
158. Laassri M, Lottenbach K, Belshe R, Wolff M, Rennels M, Plotkin S et al (2005) Effect of different vaccination schedules on excretion of oral poliovirus vaccine strains. *J Infect Dis* 192(12):2092–2098
159. Swartz TA, Green MS, Handscher R, Sofer D, Cohen-Dar M, Shohat T et al (2008) Intestinal immunity following a combined enhanced inactivated polio vaccine/oral polio vaccine programme in Israel. *Vaccine* 26(8):1083–1090
160. Lasch EE, Livni E, Englander T, El-Massri M, Marcus O, Joshua H (1978) The cell mediated immune response in acute poliomyelitis and its use in early diagnosis. *Dev Biol Stand* 41:179–182
161. Samuel BU, Cherian T, Sridharan G, Mukundan P, John TJ (1991) Immune response to intradermally injected inactivated poliovirus vaccine. *Lancet* 338(8763):343–344
162. Katrak K, Mahon BP, Minor PD, Mills KH (1991) Cellular and humoral immune responses to poliovirus in mice: a role for helper T cells in heterotypic immunity to poliovirus. *J Gen Virol* 72(Pt 5):1093–1098
163. Minor PD (2004) Polio eradication, cessation of vaccination and re-emergence of disease. *Nat Rev Microbiol* 2(6):473–482
164. Plotkin SA, Vidor E (2008) Poliovirus vaccine - inactivated. In: Plotkin SA, Orenstein WA, Offit PA (eds) *Vaccines*, 5th edn. W.B. Saunders/Elsevier, Philadelphia, pp 605–629
165. Paul J (1971) *A history of poliomyelitis*. Yale University Press, New Haven
166. Griffiths E, Wood D, Barreto L (2006) Polio vaccine: the first 50 years and beyond. *Biologicals* 34(2):73–74
167. Nathanson N, Langmuir AD (1963) The cutter incident. Poliomyelitis following formaldehyde- inactivated poliovirus vaccination in the United States during the Spring of 1955. I. Background. *Am J Hyg* 78:16–28
168. Nathanson N, Langmuir AD (1963) The cutter incident. Poliomyelitis following formaldehyde- inactivated poliovirus vaccination in the United States during the Spring of 1955. II. Relationship of poliomyelitis to cutter vaccine. *Am J Hyg* 78:29–60
169. Eddy BE, Borman GS, Berkeley WH, Young RD (1961) Tumors induced in hamsters by injection of rhesus monkey kidney cell extracts. *Proc Soc Exp Biol Med* 107:191–197
170. Shah K, Nathanson N (1976) Human exposure to SV40: review and comment. *Am J Epidemiol* 103(1):1–12

171. Mortimer EA Jr, Lepow ML, Gold E, Robbins FC, Burton GJ, Fraumeni JF Jr (1981) Long-term follow-up of persons inadvertently inoculated with SV40 as neonates. *N Engl J Med* 305(25):1517–1518
172. Swartz TA (2008) The epidemiology of polio in Israel an historical perspective. Dyonon, Tel Aviv
173. Tulchinsky T, Abed Y, Handsher R, Toubassi N, Acker C, Melnick J (1994) Successful control of poliomyelitis by a combined OPV/IPV polio vaccine program in the West Bank and Gaza, 1978–93. *Isr J Med Sci* 30(5–6):489–494
174. Tulchinsky TH, Goldblum N (2001) Polio immunization. *N Engl J Med* 344(1):61–62
175. Slater PE, Orenstein WA, Morag A, Avni A, Handsher R, Green MS et al (1990) Poliomyelitis outbreak in Israel in 1988: a report with two commentaries. *Lancet* 335(8699):1192–1195
176. Kawamura N, Kohara M, Abe S, Komatsu T, Tago K, Arita M et al (1989) Determinants in the 5' noncoding region of poliovirus Sabin 1 RNA that influence the attenuation phenotype. *J Virol* 63(3):1302–1309
177. Nomoto A, Omata T, Toyoda H, Kuge S, Horie H, Kataoka Y et al (1982) Complete nucleotide sequence of the attenuated poliovirus Sabin 1 strain genome. *Proc Natl Acad Sci USA* 79(19):5793–5797
178. Christodoulou C, Colbere-Garapin F, Macadam A, Taffs LF, Marsden S, Minor P et al (1990) Mapping of mutations associated with neurovirulence in monkeys infected with Sabin 1 poliovirus revertants selected at high temperature. *J Virol* 64(10):4922–4929
179. Macadam AJ, Pollard SR, Ferguson G, Skuce R, Wood D, Almond JW et al (1993) Genetic basis of attenuation of the Sabin type 2 vaccine strain of poliovirus in primates. *Virology* 192(1):18–26
180. Ren RB, Moss EG, Racaniello VR (1991) Identification of two determinants that attenuate vaccine-related type 2 poliovirus. *J Virol* 65(3):1377–1382
181. Macadam AJ, Arnold C, Howlett J, John A, Marsden S, Taffs F et al (1989) Reversion of the attenuated and temperature-sensitive phenotypes of the Sabin type 3 strain of poliovirus in vaccinees. *Virology* 172(2):408–414
182. Westrop GD, Wareham KA, Evans DM, Dunn G, Minor PD, Magrath DI et al (1989) Genetic basis of attenuation of the Sabin type 3 oral poliovirus vaccine. *J Virol* 63(3):1338–1344
183. Minor PD, Macadam AJ, Stone DM, Almond JW (1993) Genetic basis of attenuation of the Sabin oral poliovirus vaccines. *Biologicals* 21(4):357–363
184. Okonko IO, Babalola ET, Adedeji AO, Onoja BA, Ogun AA, Nkang AO et al (2008) The role of vaccine derived polioviruses in the global eradication of polio—the Nigeria experience as a case study. *Biotechnol Mol Biol Rev* 3(6):135–147
185. Sutter RW, Cochi SL, Melnick JL (1999) Live attenuated poliovirus vaccines. In: Plotkin S, Orenstein WA (eds) *Vaccines*, 3rd edn. W.B. Saunders, Philadelphia, pp 364–408
186. Mmwr T (2001) Apparent global interruption of wild poliovirus type 2 transmission. *MMWR Morb Mortal Wkly Rep* 50(12):222–224
187. Grassly NC, Wenger J, Durrani S, Bahl S, Deshpande JM, Sutter RW et al (2007) Protective efficacy of a monovalent oral type 1 poliovirus vaccine: a case-control study. *Lancet* 369(9570):1356–1362
188. Wkly Epidemiol Rec (2009) Advisory committee on poliomyelitis eradication: recommendations on the use of bivalent oral poliovirus vaccine types 1 and 3. *Wkly Epidemiol Rec* 84(29):289–290
189. Salk JE, Krech U, Youngner JS, Bennett BL, Lewis LJ, Bazeley PL (1954) Formaldehyde treatment and safety testing of experimental poliomyelitis vaccines. *Am J Public Health Nations Health* 44(5):563–570
190. Jenkins HE, Aylward RB, Gasasira A, Donnelly CA, Abanida EA, Koleosho-Adelekan T et al (2008) Effectiveness of immunization against paralytic poliomyelitis in Nigeria. *N Engl J Med* 359(16):1666–1674
191. Wright PF, Modlin JF (2008) The demise and rebirth of polio—a modern phoenix? *J Infect Dis* 197(3):335–336
192. Minor PD, Schild GC, Ferguson M, Mackay A, Magrath DI, John A et al (1982) Genetic and antigenic variation in type 3 polioviruses: characterization of strains by monoclonal antibodies and T1 oligonucleotide mapping. *J Gen Virol* 61(Pt 2):167–176
193. Simoes EA, Padmini B, Steinhoff MC, Jadhav M, John TJ (1985) Antibody response of infants to two doses of inactivated poliovirus vaccine of enhanced potency. *Am J Dis Child* 139(10):977–980
194. Samuel BU, Cherian T, Rajasingh J, Raghupathy P, John TJ (1992) Immune response of infants to inactivated poliovirus vaccine injected intradermally. *Vaccine* 10(2):135
195. Wkly Epidemiol Record (2004) Conclusions and recommendations of the Ad Hoc Advisory Committee on Poliomyelitis Eradication, Geneva, 21–22 September 2004. *Wkly Epidemiol Rec* 79(41):401–407
196. De Palma AM, Purstinger G, Wimmer E, Patick AK, Andries K, Rombaut B et al (2008) Potential use of antiviral agents in polio eradication. *Emerg Infect Dis* 14(4):545–551
197. Pevear DC, Tull TM, Seipel ME, Groarke JM (1999) Activity of pleconaril against enteroviruses. *Antimicrob Agents Chemother* 43(9):2109–2115
198. Oberste MS, Moore D, Anderson B, Pallansch MA, Pevear DC, Collett MS (2009) In vitro antiviral activity of V-073 against polioviruses. *Antimicrob Agents Chemother* 53(10):4501–4503
199. Levy AH (1962) The uses of gamma globulins in the prophylaxis of infection. *J Chronic Dis* 15:589–598
200. McKinney RE Jr, Katz SL, Wilfert CM (1987) Chronic enteroviral meningoencephalitis in agammaglobulinemic patients. *Rev Infect Dis* 9(2):334–356
201. Breman JG, Arita I (1980) The confirmation and maintenance of smallpox eradication. *N Engl J Med* 303(22):1263–1273
202. Smith J, Leke R, Adams A, Tangermann RH (2004) Certification of polio eradication: process and lessons learned. *Bull World Health Organ* 82(1):24–30
203. Hull HF, Ward NA, Hull BP, Milstien JB, de Quadros C (1994) Paralytic poliomyelitis: seasoned strategies, disappearing disease. *Lancet* 343(8909):1331–1337



204. Dove AW, Racaniello VR (1997) The polio eradication effort: should vaccine eradication be next? *Science* 277(5327):779–780
205. Hull HF, Aylward RB (1997) Ending polio immunization. *Science* 277(5327):780
206. Wood DJ, Sutter RW, Dowdle WR (2000) Stopping poliovirus vaccination after eradication: issues and challenges. *Bull World Health Organ* 78(3):347–357
207. Wright PF, Kim-Farley RJ, de Quadros CA, Robertson SE, Scott RM, Ward NA et al (1991) Strategies for the global eradication of poliomyelitis by the year 2000. *N Engl J Med* 325(25):1774–1779
208. WHO (1996) Field guide for supplementary activities aimed at achieving polio eradication, 1996 Revision: WHO/EPI/GEN/95.01 Rev.1
209. Hull BP, Dowdle WR (1997) Poliovirus surveillance: building the global polio laboratory network. *J Infect Dis* 175(Suppl 1):S113–S116
210. *Wkly Epidemiol Record* (2002) Expanding contributions of the global laboratory network for poliomyelitis eradication, 2000–2001. *Wkly Epidemiol Rec* 77(17):133–137
211. *Wkly Epidemiol Record* (2003) Laboratory surveillance for wild and vaccine-derived polioviruses, January 2002–June 2003. *Wkly Epidemiol Rec* 78(39):341–346
212. *Wkly Epidemiol Record* (2004) Laboratory surveillance for wild and vaccine-derived polioviruses, January 2003–June 2004. *Wkly Epidemiol Rec* 79(44):393–398
213. de Gourville E, Duintjer Tebbens RJ, Sangruejee N, Pallansch MA, Thompson KM (2006) Global surveillance and the value of information: the case of the global polio laboratory network. *Risk Anal* 26(6):1557–1569
214. van der Avoort HG, Hull BP, Hovi T, Pallansch MA, Kew OM, Crainic R et al (1995) Comparative study of five methods for intratypic differentiation of polioviruses. *J Clin Microbiol* 33(10):2562–2566
215. De L, Nottay B, Yang CF, Holloway BP, Pallansch M, Kew O (1995) Identification of vaccine-related polioviruses by hybridization with specific RNA probes. *J Clin Microbiol* 33(3):562–571
216. Kilpatrick DR, Nottay B, Yang CF, Yang SJ, Mulders MN, Holloway BP et al (1996) Group-specific identification of polioviruses by PCR using primers containing mixed-base or deoxyinosine residue at positions of codon degeneracy. *J Clin Microbiol* 34(12):2990–2996
217. Balanant J, Guillot S, Candrea A, Delpeyroux F, Crainic R (1991) The natural genomic variability of poliovirus analyzed by a restriction fragment length polymorphism assay. *Virology* 184(2):645–654
218. Kilpatrick DR, Yang CF, Ching K, Vincent A, Iber J, Campagnoli R et al (2009) Rapid group-, serotype-, and vaccine strain-specific identification of poliovirus isolates by real-time reverse transcription-PCR using degenerate primers and probes containing deoxyinosine residues. *J Clin Microbiol* 47(6):1939–1941
219. WHO (2004) Polio laboratory manual 4th edn, 2004. WHO/IVB/04.10 (database on the internet)
220. *Wkly Epidemiol Record* (2006) Resurgence of wild poliovirus type 1 transmission and effect of importation into polio-free countries, 2002–2005. *Wkly Epidemiol Rec* 81(7):63–68
221. *Wkly Epidemiol Record* (2010) Poliomyelitis in Tajikistan - first importation since Europe certified polio-free. *Wkly Epidemiol Rec* 85(18):157–158
222. Deshpande JM, Shetty SJ, Siddiqui ZA (2003) Environmental surveillance system to track wild poliovirus transmission. *Appl Environ Microbiol* 69(5):2919–2927
223. WHO (2003) Guidelines for environmental surveillance of poliovirus circulation (database on the internet). WHO, Dept of Vaccines and Biologicals; <http://www.who.int/vaccines-documents/DoxGen/H5-Surv.htm>. Available from: <http://www.who.int/vaccines-documents/DoxGen/H5-Surv.htm>
224. Manor Y, Handsher R, Halmut T, Neuman M, Abramovitz B, Mates A et al (1999) A double-selective tissue culture system for isolation of wild-type poliovirus from sewage applied in a long-term environmental surveillance. *Appl Environ Microbiol* 65(4):1794–1797
225. Shulman LM, Manor Y, Handsher R, Delpeyroux F, McDonough MJ, Halmut T et al (2000) Molecular and antigenic characterization of a highly evolved derivative of the type 2 oral poliovaccine strain isolated from sewage in Israel. *J Clin Microbiol* 38(10):3729–3734
226. Nakano JH, Hatch MH, Thieme ML, Nottay B (1978) Parameters for differentiating vaccine-derived and wild poliovirus strains. *Prog Med Virol* 24:178–206
227. Vinje J, Gregoricus N, Martin J, Gary HE Jr, Caceres VM, Venczel L et al (2004) Isolation and characterization of circulating type 1 vaccine-derived poliovirus from sewage and stream waters in Hispaniola. *J Infect Dis* 189(7):1168–1175
228. Manor Y, Handsher R, Halmut T, Neuman M, Bobrov A, Rudich H et al (1999) Detection of poliovirus circulation by environmental surveillance in the absence of clinical cases in Israel and the Palestinian authority. *J Clin Microbiol* 37(6):1670–1675
229. Ranta J, Hovi T, Arjas E (2001) Poliovirus surveillance by examining sewage water specimens: studies on detection probability using simulation models. *Risk Anal* 21(6):1087–1096
230. WHO (2004) Immunization, vaccines and biologicals. Polio laboratory manual, 4th edn. World Health Organization, Geneva, WHO/IVB/04.10
231. Hovi T, Stenvik M, Partanen H, Kangas A (2001) Poliovirus surveillance by examining sewage specimens. Quantitative recovery of virus after introduction into sewerage at remote upstream location. *Epidemiol Infect* 127(1):101–106
232. Hovi T (2006) Surveillance for polioviruses. *Biologicals* 34(2):123–126
233. Roivainen M, Blomqvist S, Al-Hello H, Paananen A, Delpeyroux F, Kuusi M et al (2010) Highly divergent neurovirulent vaccine-derived polioviruses of all three serotypes are recurrently detected in Finnish sewage. *Euro Surveill* 15(19):pii/19566
234. *Wkly Epidemiol Record* (2008) Conclusions and recommendations of the Advisory Committee on Poliomyelitis

- Eradication, Geneva, 27–28 November 2007. *Wkly Epidemiol Rec* 83(3):25–35
235. Kapp C (2003) Surge in polio spreads alarm in northern Nigeria. Rumors about vaccine safety in Muslim-run states threaten WHO's eradication programme. *Lancet* 362(9396):1631–1632
  236. Samba E, Nkrumah F, Leke R (2004) Getting polio eradication back on track in Nigeria. *N Engl J Med* 350(7):645–646
  237. *Wkly Epidemiol Record* (2009) Advisory Committee on Poliomyelitis Eradication: recommendations on the use of bivalent oral poliovirus vaccine types 1 and 3. *Wkly Epidemiol Rec* 84(29):289–290
  238. John TJ (1972) Problems with oral poliovaccine in India. *Indian Pediatr* 9(5):252–256
  239. John TJ (1976) Antibody response of infants in tropics to five doses of oral polio vaccine. *Br Med J* 1(6013):812
  240. Grassly NC, Fraser C, Wenger J, Deshpande JM, Sutter RW, Heymann DL et al (2006) New strategies for the elimination of polio from India. *Science* 314(5802):1150–1153
  241. Grassly NC, Jafari H, Bahl S, Durrani S, Wenger J, Sutter RW et al (2009) Mucosal immunity after vaccination with monovalent and trivalent oral poliovirus vaccine in India. *J Infect Dis* 200(5):794–801
  242. Gary HE Jr, Smith B, Jenks J, Ruiz J, Sessions W, Vinje J et al (2008) Failure to detect infection by oral polio vaccine virus following natural exposure among inactivated polio vaccine recipients. *Epidemiol Infect* 136(2):180–183
  243. Wringe A, Fine PE, Sutter RW, Kew OM (2008) Estimating the extent of vaccine-derived poliovirus infection. *PLoS One* 3(10):e3433
  244. *Mmwr* T (2007) Update on vaccine-derived polioviruses—worldwide, January 2006–August 2007. *MMWR Morb Mortal Wkly Rep* 56(38):996–1001
  245. *MMWR* (2009) Update on vaccine-derived polioviruses – worldwide, January 2008–June 2009. *MMWR Morb Mortal Wkly Rep* 58(36):1002–1006
  246. Dowdle W, Kew O (2006) Vaccine-derived polioviruses: is it time to stop using the word “rare?”. *J Infect Dis* 194(5):539–541
  247. Halsey NA, Pinto J, Espinosa-Rosales F, Faure-Fontenla MA, da Silva E, Khan AJ et al (2004) Search for poliovirus carriers among people with primary immune deficiency diseases in the United States, Mexico, Brazil, and the United Kingdom. *Bull World Health Organ* 82(1):3–8
  248. *Wkly Epidemiol Record* (2010) Progress toward interrupting wild poliovirus transmission worldwide. 2009. *Wkly Epidemiol Rec* 85(18):178–184
  249. Hovi T, Lindholm N, Savolainen C, Stenvik M, Burns C (2004) Evolution of wild-type 1 poliovirus in two healthy siblings excreting the virus over a period of 6 months. *J Gen Virol* 85(Pt 2):369–377
  250. Pelletier I, Duncan G, Pavio N, Colbere-Garapin F (1998) Molecular mechanisms of poliovirus persistence: key role of capsid determinants during the establishment phase. *Cell Mol Life Sci* 54(12):1385–1402
  251. Roivainen M, Blomqvist S, Al-Hello H, Paananen A, Delpeyroux F, Kuusi M et al (2010) Highly divergent neurovirulent vaccine-derived polioviruses of all three serotypes are recurrently detected in Finnish sewage. *Euro Surveill* 15(19):pii/19566
  252. ECDC (2009) Risk assessment from the ECDC on the finding of vaccine-derived polio virus in Finland February 17, 2009. \Documents and Settings\daah\Local Settings\Temporary Internet Files\OLK35A\Risk assessment polio 2009\_02\_17-JG (2).doc
  253. Pavlov DN (2006) Poliovirus vaccine strains in sewage and river water in South Africa. *Can J Microbiol* 52(8):717–723
  254. Yoshida H, Horie H, Matsuura K, Miyamura T (2000) Characterisation of vaccine-derived polioviruses isolated from sewage and river water in Japan. *Lancet* 356(9240):1461–1463
  255. Paximadi E, Karakasiliotis I, Papaventsis D, Papageorgiou G, Markoulatos P (2008) Recombinant Sabin environmental isolates in Greece and Cyprus. *J Appl Microbiol* 104(4):1153–1162
  256. CDC (2009) Update on vaccine-derived polioviruses – worldwide, January 2008–June 2009. *MMWR Morb Mortal Wkly Rep* 58(36):1002–1006
  257. Arya SC, Agarwal N (2007) Global polio laboratory network: future pursuit and commitments. *J Clin Virol* 38(4):362–363
  258. Madero E, Slacikova M, Cernakova B, Sobotova Z, Nadova K (2005) First isolation of vaccine-derived poliovirus in Slovakia. *Euro Surveill* 10(8):E050818 3
  259. Alexander JP, Ehresmann K, Seward J, Wax G, Harriman K, Fuller S et al (2009) Transmission of imported vaccine-derived poliovirus in an undervaccinated community in Minnesota. *J Infect Dis* 199(3):391–397
  260. *MMWR* (1994) Certification of poliomyelitis eradication – the Americas, 1994. *MMWR Morb Mortal Wkly Rep* 43(39):720–722
  261. Strebel PM, Sutter RW, Cochi SL, Biellik RJ, Brink EW, Kew OM et al (1992) Epidemiology of poliomyelitis in the United States one decade after the last reported case of indigenous wild virus-associated disease. *Clin Infect Dis* 14(2):568–579
  262. *MMWR* (2001) Erratum: apparent global interruption of wild poliovirus type 2 transmission. *MMWR* 50(12):249
  263. *MMWR* (2001) Certification of poliomyelitis eradication—Western Pacific Region, October 2000. *MMWR Morb Mortal Wkly Rep* 50(1):1–3
  264. *Wkly Epidemiol Record* (2005) Conclusions and recommendations of the Advisory Committee on Poliomyelitis Eradication, Geneva, 11–12 Oct 2005. *Wkly Epidemiol Rec* 80(47):410–416
  265. Heymann DL, Sutter RW, Aylward RB (2006) A vision of a world without polio: the OPV cessation strategy. *Biologicals* 34(2):75–79
  266. Ehrenfeld E, Modlin J, Chumakov K (2009) Future of polio vaccines. *Expert Rev Vaccines* 8(7):899–905
  267. Jenkins PC, Modlin JF (2006) Decision analysis in planning for a polio outbreak in the United States. *Pediatrics* 118(2):611–618
  268. Tebbens RJ, Pallansch MA, Alexander JP, Thompson KM (2010) Optimal vaccine stockpile design for an eradicated disease: application to polio. *Vaccine* 28(26):4312–4327

269. WHO (2005) WHO framework for national policy makers in OPV-using countries – Cessation of routine oral polio vaccine (OPV) use after global polio eradication 05.02.: WHO/POLIO/05.02
270. Wagner BG, Earn DJ (2008) Circulating vaccine derived polio viruses and their impact on global polio eradication. *Bull Math Biol* 70(1):253–280
271. Fine PE, Sutter RW, Orenstein WA (2001) Stopping a polio outbreak in the post-eradication era. *Dev Biol (Basel)* 105:129–147
272. Tebbens RJ, Pallansch MA, Kew OM, Caceres VM, Jafari H, Cochi SL et al (2006) Risks of paralytic disease due to wild or vaccine-derived poliovirus after eradication. *Risk Anal* 26(6):1471–1505
273. Oliver Wyman Inc (2010) The supply landscape and economics of IPV-containing combination vaccines: Key findings. May 2010: Commissioned by the Bill & Melinda Gates Foundation
274. Chumakov K, Ehrenfeld E (2008) New generation of inactivated poliovirus vaccines for universal immunization after eradication of poliomyelitis. *Clin Infect Dis* 47(12):1587–1592
275. Baca-Estrada M, Griffiths E (2006) Regulation and standardization of IPV and IPV combination vaccines. *Biologicals* 34(2):159–161
276. Duchene M (2006) Production, testing and perspectives of IPV and IPV combination vaccines: GSK biologicals' view. *Biologicals* 34(2):163–166
277. Graf H (2006) Manufacturing and supply of monovalent oral polio vaccines. *Biologicals* 34(2):141–144
278. El-Sayed N, El-Gamal Y, Abbassy AA, Seoud I, Salama M, Kandeel A et al (2008) Monovalent type 1 oral poliovirus vaccine in newborns. *N Engl J Med* 359(16):1655–1665
279. Farag MM (2006) Licensing of monovalent OPV1 vaccine. *Biologicals* 34(2):145–149
280. Caceres VM, Sutter RW (2001) Sabin monovalent oral polio vaccines: review of past experiences and their potential use after polio eradication. *Clin Infect Dis* 33(4):531–541
281. Plotkin JB, Dushoff J (2003) Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. *Proc Natl Acad Sci USA* 100(12):7152–7157
282. Macadam AJ, Ferguson G, Stone DM, Meredith J, Knowlson S, Auda G et al (2006) Rational design of genetically stable, live-attenuated poliovirus vaccines of all three serotypes: relevance to poliomyelitis eradication. *J Virol* 80(17):8653–8663
283. Eichner M, Dietz K (1996) Eradication of poliomyelitis: when can one be sure that polio virus transmission has been terminated? *Am J Epidemiol* 143(8):816–822
284. Vashishtha VM (2004) But do we have other options? *Indian J Pediatr* 71(2):183–184
285. Yang CF, Naguib T, Yang SJ, Nasr E, Jorba J, Ahmed N et al (2003) Circulation of endemic type 2 vaccine-derived poliovirus in Egypt from 1983 to 1993. *J Virol* 77(15):8366–8377
286. Shulman LM, Manor J, Sofer D, Mendelsohn E (2009) Environmental surveillance for polioviruses in Israel: Bio-error, Bio-terror or just mother nature. In: Marks RS, Lobel L, Amadou SA (eds) *Advanced detection of viral pathogens*. Neobionics, Omer, Israel, pp 111–121
287. Nathanson N (2008) The pathogenesis of poliomyelitis: what we don't know. *Adv Virus Res* 71:1–50
288. Minor P (2006) Gaps in scientific knowledge for the post eradication world. *Biologicals* 34(2):167–170
289. Willerth SM, Pedro HA, Pachter L, Humeau LM, Arkin AP, Schaffer DV (2010) Development of a low bias method for characterizing viral populations using next generation sequencing technology. *PLoS One* 5(10):e13564
290. Kitamura N, Semler BL, Rothberg PG, Larsen GR, Adler CJ, Dorner AJ et al (1981) Primary structure, gene organization and polypeptide expression of poliovirus RNA. *Nature* 291(5816):547–553
291. Shulman LM, Manor Y, Sofer D, Mendelson E (2012) Environmental surveillance for poliovirus in Israel. In: Marks RS (eds) *Bio-error, Bio-terror, or just mother nature*. In viral detection biosensors, Pan Stanford Publishing, Singapore (In press)

## Polybenzimidazole Fuel Cell Technology

MAX MOLLE<sup>1</sup>, THOMAS J. SCHMIDT<sup>2</sup>,  
BRIAN C. BENICEWICZ<sup>1</sup>

<sup>1</sup>Department of Chemistry and Biochemistry,  
University of South Carolina, Columbia, SC, USA

<sup>2</sup>General Energy Research, Lab. of Electrochemistry,  
Paul Scherrer Institute, Villigen-PSI, Switzerland

### Article Outline

Glossary

Definition of the Subject

Introduction to Polybenzimidazole Fuel Cell

Sustainability

History and Technical Information of Polybenzimidazole Membranes

PBI-PA Fuel Cell Systems and Their Applications

Conclusions and Future Directions

Acknowledgments

Bibliography

### Glossary

**Polybenzimidazoles (PBIs)** A class of polymers recognized for their excellent thermal and chemical stability, PBIs have historically been spun into fibers and woven into thermal protective clothing. In the past decade, PBIs have been cast into membranes and incorporated into fuel cells.

**Polymer electrolyte membrane (PEM)** Also referred to as Proton Exchange Membranes, PEMs are semi-permeable membranes that conduct and transport protons while preventing the transmission of gases and electrons.

**Membrane electrode assembly (MEA)** A device that is comprised of a PEM that is sandwiched between two electrodes.

**Conventional imbibing** The original process of impregnating polymer membranes with dopants. The precast, fully dense membranes are placed in baths of dopants and allowed to absorb the dopant which assists in proton conductivity.

**PPA process** A recently developed imbibing process, PBIs are polymerized and cast in a polyphosphoric acid (PPA) solvent. Under controlled hydrolysis conditions, Polyphosphoric acid, a good solvent for PBI, is converted into phosphoric acid, a poor solvent for PBI. A mechanically stable PBI gel membrane that is highly doped with phosphoric acid is produced by means of a sol-to-gel transition.

**Proton conductivity** A measure of how well a material can transfer protons. In fuel cell technology, it is used to gauge the viability of proton exchange membranes.

**Combined heat and power (CHP)** Stationary fuel cell devices that are used to produce both heat and electricity. High-temperature PBI fuel cell membranes are well suited for this application.

## Definition of the Subject

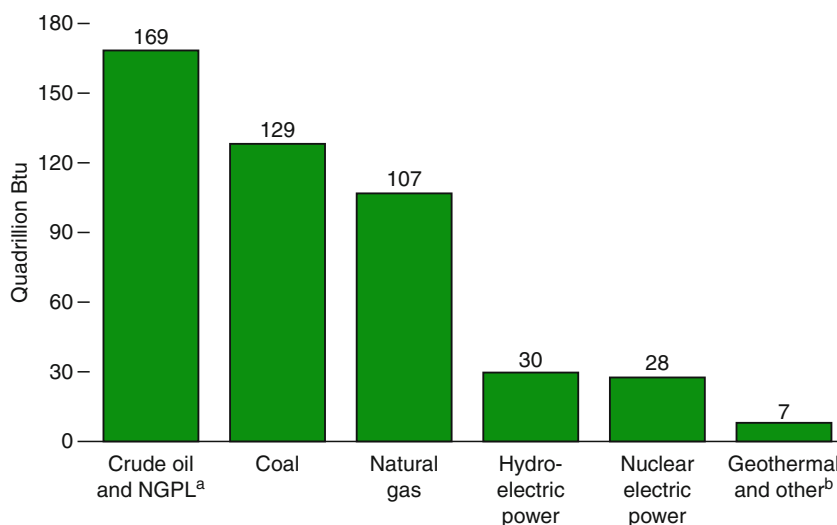
After approximately 10 years of development, polybenzimidazole (PBI) chemistries and the concomitant manufacturing processes have evolved into commercially produced membrane electrode assemblies (MEAs). PBI MEAs can operate reliably without complex water humidification hardware and are able to run at elevated temperatures of 120–180°C due to the physical and chemical robustness of PBI membranes. These higher temperatures improve the electrode kinetics and conductivity of the MEAs, simplify the water and thermal management of the systems, and significantly increase their tolerance to fuel impurities. Membranes cast by a newly developed polyphosphoric acid (PPA) process possessed excellent mechanical properties, higher phosphoric acid (PA)/PBI ratios, and enhanced

proton conductivities as compared to previous methods of membrane preparation. *p*-PBI is the most common polymer in PBI-based fuel cell systems, although AB-PBI and other derivatives have been investigated. This chapter reports on the chemistries and sustainable usages of PBI-based high-temperature proton exchange membrane fuel cells (PEMFCs).

## Introduction to Polybenzimidazole Fuel Cell Sustainability

Alternative energy is often defined as any energy derived from sources other than fossil fuels or nuclear fission. These alternative energy sources, which include solar, wind, hydro, and geothermal energy, are considered renewable because they are naturally replenished and their supply is seemingly limitless. In contrast, the Earth's supply of fossil fuels is constantly being diminished. Fossil fuels, which include crude oil, coal, and natural gas, continue to be the dominating sources of energy in the world (Fig. 1). Fossil fuels provide more than 86% of the total energy consumed globally [1]. In 2008, over two-thirds of the electrical energy and 97% of the transportation energy in the USA was produced from these nonrenewable sources [2]. It is predicted that the global demand for fossil fuels will continue to increase over the next 10–20 years due to economic growth. One may conclude that the importance of renewable energy will steadily increase as the Earth's supply of fossil fuels continues to be depleted.

Polymer electrolyte membrane (PEM) fuel cells, also known as proton exchange membrane fuel cells (PEMFCs), are energy conversion devices that could provide the world with clean and efficient energy. Due to their excellent energy production, inexpensive starting materials, and lack of pollutant by-products, these cells have exponentially gained in popularity over the past decade. Electricity is produced at the heart of the fuel cell by the membrane electrode assembly (MEA), a component that is comprised of a proton exchange membrane sandwiched between two electrodes. Fueled by a hydrogen-based source, a metal catalyst at the anode splits the hydrogen into protons and electrons. As the protons are transported through the proton electrolyte membrane to the cathode, the electrons provide electrical work by traveling around the membrane through an external circuit from the

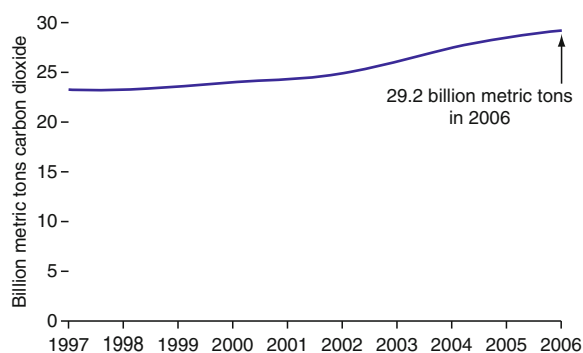


**Polybenzimidazole Fuel Cell Technology. Figure 1**

Global production of energy in 2006 by source. <sup>a</sup>Natural gas plant liquid. <sup>b</sup>Net electricity generation from wood, waste, solar, and wind [1]

anode to the cathode. The protons and electrons react with an oxidant (typically air or pure oxygen) at the cathode to form water, thereby completing the electrochemical cycle. Hydrogen gas is commonly used as a fuel source for the cells, but other fuels such as methane, methanol, and ethanol have been explored.

PEM fuel cells provide multiple advantages over conventional fossil fuel energy production. Because water is the only by-product of the electrochemical process, these fuel cells are clean and environmentally friendly. If one considers the tremendous amount of carbon dioxide created by energy production on the global scale (Fig. 2), PEM fuel cells offer a method to significantly reduce hazardous gas emissions. Minimal moving parts reduces the amount of maintenance of each cell, and the lack of combustion significantly decreases the amount of harmful pollutants such as sulfur oxides and nitrogen oxides. In addition, PEM fuel cells are much more efficient at producing energy (this is discussed in detail in section “PBI-PA Fuel Cell Systems and their Applications”), and much like a combustion engine, the cell can run continuously as long as fuel and oxidant are provided. Although fuel cells are an environmentally friendly energy conversion device, one must consider the manner in which hydrogen is gathered. Both hydrogen production and conversion from chemical to electrical energy need to be



**Polybenzimidazole Fuel Cell Technology. Figure 2**

Global production of carbon dioxide annually from 1997 to 2006 [3]

sustainable to make the overall process sustainable. Hydrogen production, however, is out of the scope of this chapter.

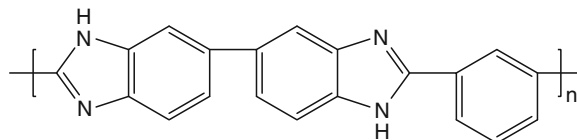
The efficiency of a PEM fuel cell is largely dependent on the materials used and their arrangement in the cell. Fuel cells use an array of different catalysts, electrodes, membranes, and dopants, each of which function under specific operating conditions. Cells that use low-boiling dopants, such as water, operate at approximately 60–80°C to avoid vaporization of the proton-transfer agent. Large heat exchangers are required to

ensure the heat generated by the cell does not vaporize the electrolyte. Consequently, system complexity is increased as extra components and controls are required to ensure that the membrane remains hydrated during operation. Moreover, cell operation at such low temperatures allows trace amounts of reformat by-products, especially carbon monoxide, to bind to the catalyst. These highly competitive, nonreversible reactions “poison” the catalyst, thereby decreasing and possibly terminating the functionality of the fuel cell. Therefore, low-temperature fuel cells require an extremely pure fuel source.

In contrast to low-temperature cells, high-temperature PEMs use high-boiling dopants, such as phosphoric acid and sulfuric acid, and function at temperatures of 120–200°C. Because the cell is able to run at elevated temperatures, much smaller heat exchangers are required. Operating at higher temperatures allows fuel pollutants to bind reversibly to the catalyst, which helps to prevent catalyst poisoning. Comparatively, high-temperature PEMs can use reformed gases with much higher levels of impurities and lower reformation costs. Furthermore, high temperatures typically improve both the electrode kinetics and operating abilities of the cell. This chapter reports on the chemistries and sustainable usages of PBI-based high-temperature PEMFCs.

### History and Technical Information of Polybenzimidazole Membranes

Polybenzimidazoles (PBIs) are a class of polymers recognized for their excellent thermal and chemical stability. PBI is used in multiple applications including matrix resins, high-strength adhesives, thermal and electrical insulating foams, and thermally resistant fibers. PBI fibers were originally synthesized in the early 1960s by a cooperative effort of the US Air Force Materials Laboratory with Dupont and the Celanese Research Company. One of the first PBIs to be widely investigated was poly(2,2'-*m*-phenylene-5,5'-bibenzimidazole), which is commonly referred to as *m*-PBI (Fig. 3). Because *m*-PBI is nonflammable, resistant to chemicals, physically stable at high temperatures, and can be spun into fibers, this polymer has been used in astronaut space suits, firefighter's turnout coats and suits, and high-temperature protective gloves.



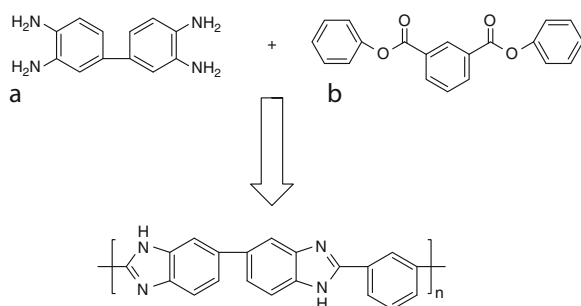
**Polybenzimidazole Fuel Cell Technology. Figure 3**  
Chemical structure of poly (2,2'-*m*-phenylene-5,5'-bibenzimidazole) (*m*-PBI)

Acid-doped polybenzimidazole membranes are excellent candidates for high-temperature fuel cells because of their thermal and chemical stability and proton conducting ability. The stability of PBIs is attributed to its aromatic structure (alternating single and double bonds) and the rigid nature of its bonds [4]. While the acid-doped membrane structure allows protons to flow from one side to the other, it acts as a barrier to the crossover of gases and electrons. The chemical stability of PBIs allows the membranes to withstand the chemically reactive environments of the anode and cathode. Furthermore, the basic nature of the polymer allows it to be highly doped with phosphoric or sulfuric acid. The dopants interact with the polymer matrix and provide a network through which protons can be transported. These acids are used as electrolytes because of their high conductivity, thermal stability, and enhanced proton-transport capabilities. It is important to note that the proton conductivity of PBI membranes without a dopant is negligible. For liquid phosphoric acid, the proton jump rate is orders of magnitude larger than the diffusion of the phosphoric acid molecule as a whole [5]. Additionally, it has been reported that both protons and phosphate moieties have a substantially decreased diffusion coefficient when blended with basic polymers as opposed to liquid phosphoric acid [6]. Therefore, a heterogeneous, two-phase system in which the PBI membrane is phase-separated and imbibed with phosphoric acid has a higher conductivity than its homogeneous counterpart [7]. The partial charges of the phosphate ions involved with proton transfers increases charge delocalization, which lowers the overall energy barrier of proton transfer [8]. The proton-transfer mechanism of large proton vehicle species (such as phosphate ions) can be initiated by local vibrations of the vehicle species. In comparison, the amount of energy required to induce proton transfer small proton vehicle species

such as water is comparable to the amount of energy required to diffuse the entire small proton vehicle.

### Synthesis of Polybenzimidazoles

One of the first PBI membranes investigated for fuel cell use was poly(2,2'-*m*-phenylene-5,5'-bibenzimidazole) (*m*-PBI). At the time, there was a vast amount of research previously reported on *m*-PBI and it was renowned for its excellent thermal and mechanical properties [5]. The polymer is synthesized by the reaction of 3,3',4,4'-tetraaminobiphenyl (TAB) with diphenylisophthalate (DPIP) during a melt/solid polymerization (Scheme 1). The resulting polymer is extracted and has an inherent viscosity (IVs) between 0.5 and 0.8 dL g<sup>-1</sup>, which corresponds to a polymer with low to moderate molecular weight. The *m*-PBI is further purified by dissolving it in a solution of N,N-dimethylacetamide and lithium chloride (DMAc/LiCl) under 60–100 psi and 250°C and then filtering; this step removes any cross-linked *m*-PBI. The polymer is then cast as a film and dried at 140°C under vacuum to evaporate the solvent. The *m*-PBI membrane is washed in boiling water to remove any residual DMAc/LiCl solution trapped in the polymer matrix. After the polymer has been dried, an acid bath is used to dope the membrane; the doping level of the membrane can be partially controlled by varying the concentration of acid in the bath. Originally, this conventionally imbibed process created membranes with molar ratios of phosphoric acid/polymer repeat unit (PA/PRU) approximately 6–10 [9]. A “direct acid casting” (DAC) technique was later developed to allow



**Polybenzimidazole Fuel Cell Technology. Scheme 1**  
Polymerization of *m*-PBI from 3,3',4,4'-tetraaminobiphenyl (a) and diphenylisophthalate (b)

the PBI membrane to retain more PA [10]. Both the conventional imbibing process and DAC were developed following the research performed by Jean-Claude Lassegues, who was one of the first scientists that investigated basic polymeric-acid systems (a summary of his work is reviewed in reference [11]). The DAC technique consists of extracting low molecular weight PBI components from PBI powder, and then dissolving the high molecular weight PBI components in trifluoroacetic acid (TFA). Phosphoric acid is added to the TFA/PBI mixture, which is then cast onto glass plates with a casting blade. One may tune the doping level of the polymer by adjusting the amount of phosphoric acid that is added to the TFA/PBI mixture. However, as one increases the PA doping level of a DAC PBI membrane, its mechanical strength decreases to the point where it can no longer be used in a fuel cell. Modern imbibing processes can increase the PA/PBI ratio to 12–16, and these fuel cell membranes are reported to have proton conductivities as high as 0.08 S cm<sup>-1</sup> at 150°C at various humidities.

A novel synthetic process for producing high molecular weight PBIs, the “PPA Process” was developed at Rensselaer Polytechnic Institute with cooperation from BASF Fuel Cell GmbH. This process has previously been discussed by Xiao et al. [12]. The general synthesis of PBI by this method requires the combination of a tetraamine with a dicarboxylic acid in polyphosphoric acid (PPA) in a dry environment. The step-growth polycondensation reaction typically occurs around 200°C for 16–24 h in a nitrogen atmosphere, producing high molecular weight polymer. This solution is cast directly from PPA as a thin film on a substrate, and upon absorption of water, the PPA hydrolyzes in situ to form phosphoric acid. Note that PPA is a good solvent for PBI while PA is a poor solvent. Under controlled hydrolysis conditions, a mechanically stable PBI gel membrane that is highly doped with phosphoric acid is produced. The multiple physical and chemical transformations that explain the solution-to-gel phase transition are summarized in Fig. 4.

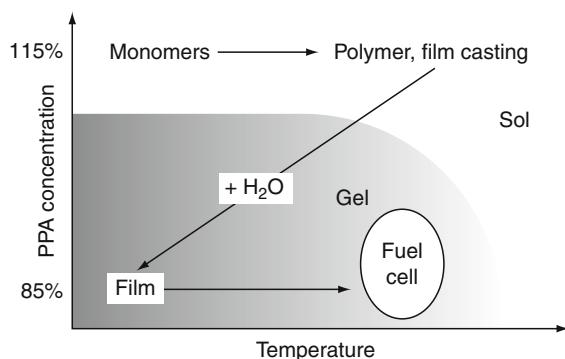
The PA-doped *m*-PBI fuel cell membrane maintains thermal and physical stability while operating at high temperature. To illuminate the fundamental differences in polymer film architecture, polymers with similar physical characteristics were prepared by the conventional PPA Process (Table 1). Even though the

ratio of phosphoric acid-to-polymer repeat unit (PA/PRU) achieved by both processes were nearly identical, the PPA Process produces membranes with much higher proton diffusion coefficients and conductivities. The higher proton diffusion coefficients of membranes produced by the PPA Process versus conventionally imbibed membranes were confirmed by NMR [13]. One can conclude that the PPA Process creates a membrane with a proton-transport architecture superior to that of the conventionally imbibed PBI membrane. In addition, inherent viscosity data indicates that the PPA process produces polymers of much higher molecular weight [12]. It was subsequently shown that improved membrane morphology and increased molecular weight allow the polymer to retain much more phosphoric acid than traditionally cast PBI membranes. An increased PA doping level typically improves the

conductivity of the membrane and may even increase the performance of the cell.

### Properties and Performance of Synthetically Modified PBI

In this chapter, the synthesis of significant PBI membranes (Fig. 5) and their use in fuel cells are described. Synthetically modified PBIs are investigated for enhanced thermo-oxidative stability, solubility, and flexibility; these attributes allow for improved processability and production of membranes with good chemical and mechanical properties. All PBI membranes are produced by means of step-growth polycondensation reactions and are generally imbibed by either the conventional technique or made by the PPA process. To synthesize modified polymers, one may either polymerize modified monomers or use post-polymerization cross-linking or substitution reactions. The following sections briefly detail the syntheses of PBI derivatives and their performances as fuel cell membranes.



**Polybenzimidazole Fuel Cell Technology. Figure 4** State diagram of the PPA Sol-gel process [12]

***m*-PBI** One of the first PBI membranes investigated for fuel cell use was *m*-PBI (Fig. 5a). As previously discussed, the film can be processed by using either the conventional imbibing method or the PPA process. Using the conventional imbibing method, the inherent viscosity of the membrane is usually between 0.50–1.00 dL g<sup>-1</sup> at 30°C, which indicates polymers of moderate molecular weight. In contrast, *m*-PBI membranes synthesized and doped via the PPA Process have inherent viscosities of approximately 1.00–2.35 dL g<sup>-1</sup> at 30°C, which corresponds to higher molecular weight polymers [9]. Using

**Polybenzimidazole Fuel Cell Technology. Table 1** Comparison of conventionally imbibed *m*-PBI versus *m*-PBI synthesized from the PPA Process [14]

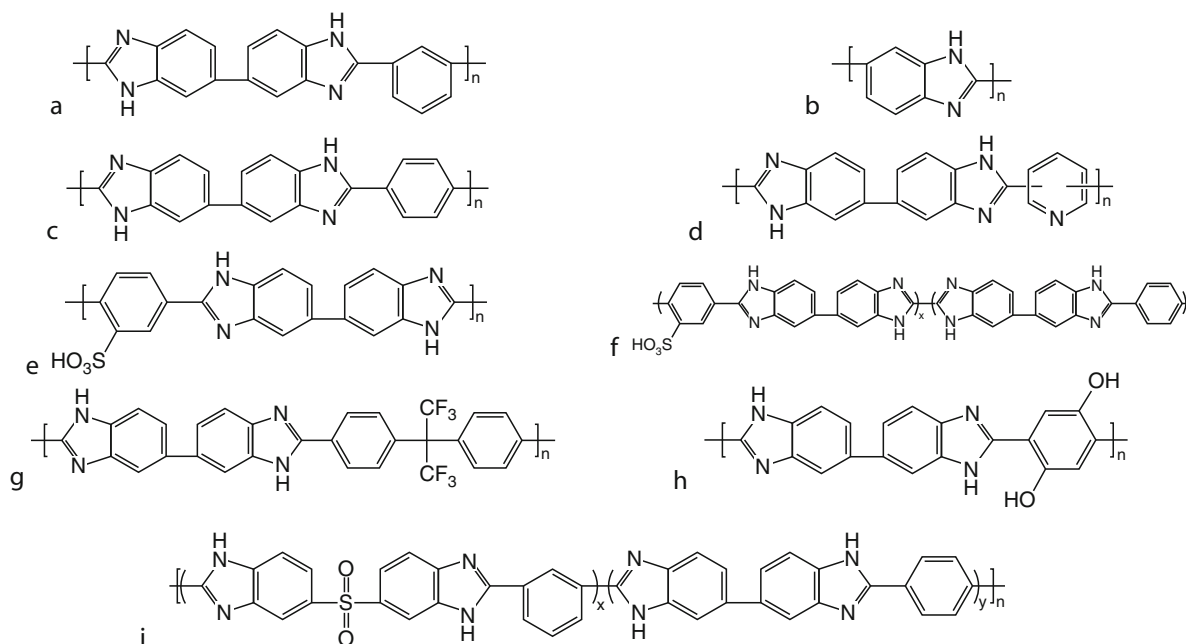
IV <sup>a</sup> (dL g <sup>-1</sup> )	Film process	Polymer (wt%)	PA (wt%)	Water (wt%)	PA/PBI (Molar ratio)	Proton diffusion coefficient <sup>b</sup> (cm <sup>2</sup> s <sup>-1</sup> )	Conductivity <sup>c</sup> (S cm <sup>-1</sup> )
0.89	Conventionally imbibed	15.6	60.7	23.7	12.2	10 <sup>-7</sup>	0.048
1.49	PPA Process	14.4	63.3	22.3	13.8	3 × 10 <sup>-6</sup>	0.13

<sup>a</sup>Inherent viscosity (IV) was measured at a polymer concentration of 0.2 g dL<sup>-1</sup> in concentrated sulfuric acid (96%) at 30°C, using a Canon Ubbelohde viscometer

<sup>b</sup>Estimation of upper bound for conventionally imbibed *m*-PBI at 180°C; PPA-prepared *m*-PBI measured at 180°C

<sup>c</sup>Measured at 160°C after an initial heating to 160°C to remove water





### Polybenzimidazole Fuel Cell Technology. Figure 5

Various synthetically modified polybenzimidazoles for use in fuel cells. (a) *m*-PBI, (b) AB-PBI, (c) *p*-PBI, (d) py-PBI, (e) *s*-PBI, (f) *s*-PBI/*p*-PBI random block copolymer, (g) 6F-PBI, (h) 2OH-PBI, and (i) *m*-SPBI/*p*-PBI segmented block copolymer

the PPA Process, higher molecular weight polymers have contributed to higher doping levels. Phosphoric acid doping levels for conventionally prepared *m*-PBI ranged from 6 to 10 moles PA/PRU, whereas the doping levels for polymer films prepared via the PPA process range from 14 to 26 moles PA/PRU [4]. Trends show that the mechanical stability of conventionally prepared membranes decrease as the doping level increases and/or as the molecular weight of the polymer decreases. The doping level, casting technique, temperature, and humidity all influence the conductivity of a *m*-PBI membrane. Under various humidities, conventionally prepared *m*-PBI membranes have been reported having conductivities in the range of 0.04–0.08 S cm<sup>-1</sup> [15]. Using the PPA Process, the conductivity values of *m*-PBI membranes are typically higher than that of the conventionally imbibed process. One study [16] reported *m*-PBI membranes formed by the PPA Process as having a conductivity of 0.13 S cm<sup>-1</sup> at 160°C under non-humidified conditions.

Phosphoric acid-doped *m*-PBI membranes that have been formed by the conventional imbibing method have been extensively studied for use in fuel

cells. Li et al. [17] demonstrated that a membrane with 6.2 PA/PRU doping level obtains a current density of approximately 0.7 A cm<sup>-2</sup> at 0.6 V using hydrogen and oxygen gases; these results were promising because the gases were not humidified. Zhai et al. [18] studied the degradation mechanisms of the PA/*m*-PBI system by continuously operating it at 0.640 A cm<sup>-2</sup> at 150°C with unhumidified hydrogen and oxygen for 550 h; the fuel cell was operated intermittently the last 50 h with shutoffs every 12 h. The voltage increased from 0.57 to 0.66 V during the beginning 90 h activation period, and the following 450 h period showed a steady decrease to 0.58 V. The performance of the system rapidly decreased in the following 10 h due to agglomeration of the platinum from the catalyst, leaching of the phosphoric acid, and hydrogen crossover. Kongstein et al. [19] employed use of a dual layer electrode to prevent the oxidation of carbon in the polymer membrane, which can occur in acidic environments at high voltages. This electrode would improve the structural integrity of the polymer and help prevent hydrogen crossover from occurring. The PA/*m*-PBI membrane had a maximum of 0.6 V at 0.6 A cm<sup>-2</sup> with a maximum power density

of  $0.83 \text{ W cm}^{-2}$  at 0.4 V. These performances were lower than that of other PEM systems, such as Nafion, but were still impressive because they could be run at much higher temperatures.

**Poly(2,5-Polybenzimidazole): AB-PBI** Commonly referred to as AB-PBI, poly(2,5-polybenzimidazole) has a much simpler structure than that of *m*-PBI and other polybenzimidazoles (Fig. 5b). Whereas *m*-PBI is synthesized from 3,3',4,4'-tetraaminobiphenyl and DPIP, AB-PBI is polymerized from a single monomer, 3,4-diaminobenzoic acid (DABA). This monomer is commercially available and is less expensive than the starting materials of *m*-PBI. The polymer membrane can be cast and imbibed with phosphoric acid by the conventional imbibing method in a mixture of methanesulfonic acid (MSA) and phosphorous pentoxide ( $\text{P}_2\text{O}_5$ ) [20] or DMAc. It can also be cast by direct acid casting using trifluoroacetic acid (TFA) [10, 15] or by the PPA Process [10, 21–23]. AB-PBI membranes prepared by the conventional imbibing method had IV values around  $2.0\text{--}2.5 \text{ dL g}^{-1}$  as reported by Asensio et al. [23] and  $6\text{--}8 \text{ dL g}^{-1}$  by Litt et al. [15]. Polymers produced from recrystallized DABA by the PPA Process have IV values greater than  $10 \text{ dL g}^{-1}$  [24]; however, membranes of AB-PBI could not be easily formed via the PPA Process because of the polymer's high solubility in acids.

Because AB-PBI has a high concentration of basic sites (amine and imine groups), it has a high solubility and affinity to acids. Due to this affinity, it can be doped with phosphoric acid and sulfonated with sulfuric acid. Sulfonation of AB-PBI (sAB-PBI) is

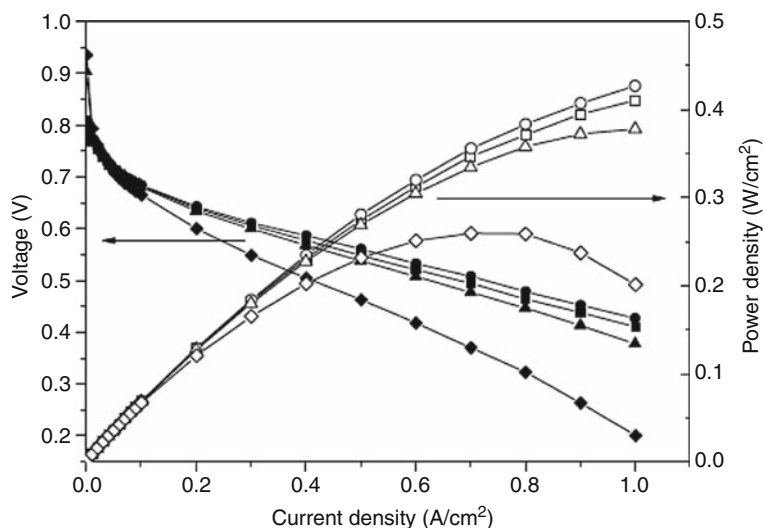
performed by soaking the precast polymer in sulfuric acid followed by treating the mixture with heat. Asensio et al. [23] reported sAB-PBI-PA membranes having an enhanced conductivity over that of AB-PBI-PA and to be both mechanically strong and thermally stable. Using the direct casting method from MSA- $\text{P}_2\text{O}_5$ , Kim et al. [20] produced AB-PBI-PA membranes with conductivities similar that of Asensio, having values ranging from  $0.02\text{--}0.06 \text{ S cm}^{-1}$  at  $110^\circ\text{C}$  with no humidification. The conductivity values and physical-chemical properties resemble that of *m*-PBI, making it a good candidate for fuel cell use.

Yu [25] synthesized *p*-PBI-block-AB-PBI membranes to lower the membrane's solubility in acids while maintaining a high acid doping level. Different molar ratios of each polymer block were synthesized, and their conductivities and acid doping levels were investigated. As detailed in Table 2, the proton conductivities of the segmented block copolymers were enhanced by an order of magnitude over that of native AB-PBI. Stress-strain studies showed that these block copolymers were strong enough to be used in fuel cell tests. Polarization curves (Fig. 6) of these membranes illustrate that copolymers II, III, and IV have excellent fuel cell properties (approximately  $0.6 \text{ V}$  at  $0.2 \text{ A cm}^{-2}$ ); polarization curves for copolymer V and VI could not be measured due to poor thermal stability of the membrane (re-dissolution) at  $160^\circ\text{C}$ .

**Poly(2,2'-(1,4-Phenylene)5,5'-Bibenzimidazole): *p*-PBI** Poly(2,2'-(1,4-phenylene)5,5'-bibenzimidazole) (*p*-PBI, Fig. 5c) is one of the highest performing PBI membranes for high-temperature fuel cell use. Due to the

**Polybenzimidazole Fuel Cell Technology. Table 2** Percent composition, acid doping level, and proton conductivity data for various *p*-PBI-block-AB-PBI membranes [25]

	<i>Para</i> -PBI/AB-PBI (mole ratio, <i>x/y</i> )	Acid doping level (PA/2 benzimidazole)	Proton conductivity (S/cm @ $160^\circ\text{C}$ )	Membrane composition (%)		
				Polymer	$\text{H}_3\text{PO}_4$	Water
I	100/0	42.9	0.25	4.13	60.38	35.11
II	75/25	19.1	0.25	8.68	58.09	33.22
III	50/50	24.1	0.27	6.59	54.53	38.88
IV	25/75	21.8	0.23	7.79	63.31	28.90
V	10/90	17.3	0.15	8.84	63.23	27.94
VI	0/100	N/A	N/A			

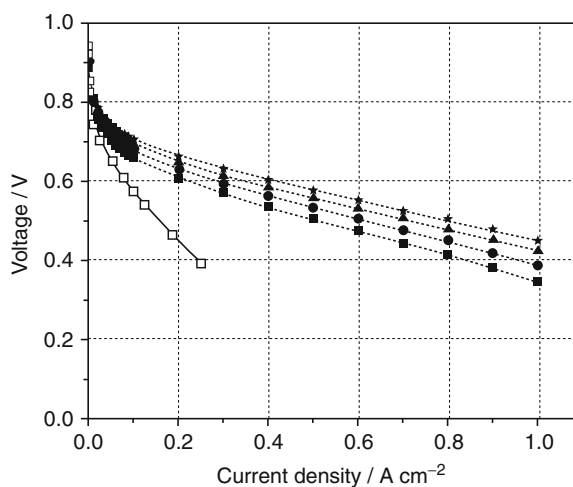


**Polybenzimidazole Fuel Cell Technology. Figure 6**

Polarization curves (filled symbols) and power density curves (unfilled symbols) of *p*-PBI (Polymer I, ■) and *p*-PBI-block-AB-PBI membranes (75/25, Polymer II, ●, 50/50 Polymer III, ▲, 25/75, Polymer IV, ◆) at 160°C with H<sub>2</sub> (1.2 stoic)/Air (2.0 stoic) under atmospheric pressure [25]

rigid nature of *p*-PBI, high molecular weight polymers have typically been difficult to fabricate or process. The first reported high molecular weight *p*-PBI with an IV value of 4.2 dL g<sup>-1</sup> was synthesized in 1975 by the US Air Force Materials Laboratory [26]. Because it could not be spun into fibers as easily as *m*-PBI, *p*-PBI was not investigated further until after the turn of the century. Using the PPA Process, Xiao et al. [12] and Yu et al. [27] synthesized high molecular weight *p*-PBI with IV values as high as 3.8 dL g<sup>-1</sup>. The PA doping level of the corresponding polymer membranes was >30 mol PA/PRU, allowing the membrane to achieve a conductivity of 0.24 S cm<sup>-1</sup> at 160°C. Xiao and Yu showed that *p*-PBI membrane achieves a much higher acid doping level and conductivity than that of *m*-PBI, which only achieves a doping level of 13–16 mol PA/PRU with a conductivity of 0.1–0.13 S cm<sup>-1</sup>. Because *p*-PBI had excellent mechanical properties at this high doping level, it was a prime candidate for fuel cell performance tests.

The polarization curves of an MEA using *p*-PBI produced by the PPA Process at various temperatures are shown in Fig. 7. Hydrogen was used as the fuel and air was used as the oxidant. The *p*-PBI outperformed the *m*-PBI at all temperatures, and the performance of the MEA increased as the temperature increased. Using



**Polybenzimidazole Fuel Cell Technology. Figure 7**

Polarization curves of PPA-Processed *p*-PBI MEA using hydrogen/air at 120°C (squares), 140°C (circles), 160°C (triangles), and 180°C (stars). Open squares represent DMAc cast *m*-PBI MEA at 150°C [27]

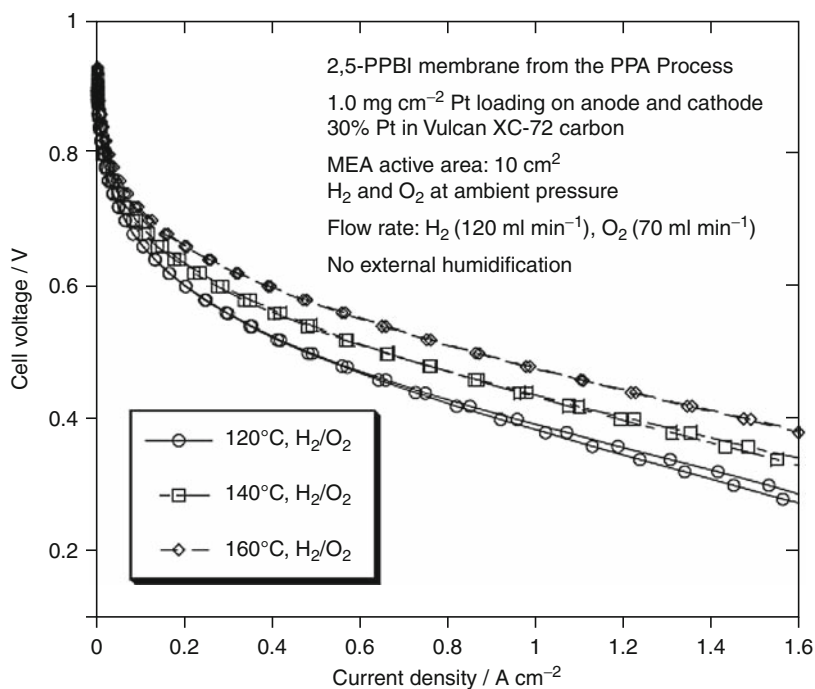
a load of 0.2 A cm<sup>-2</sup>, the cell was able to produce a voltage of 0.606 V at 120°C; upon raising the temperature to 180°C, the voltage increased to 0.663 V. This was especially promising because the gases were unhumidified.

**Pyridine-PBI** Pyridine polybenzimidazoles (py-PBIs, Fig. 5d) have been investigated for their use in fuel cells because of their high concentration of basic sites (amine and imine groups). Similar to AB-PBI, the high concentration of basic sites allow these polymers to have a high affinity to acids. The pyridine moiety is commonly combined with the traditional PBI structure by including it as part of the backbone structure.

Xiao et al. synthesized an array of py-PBIs that have the pyridine moiety as part of the polymer backbone [12, 28, 29]. These polymers were synthesized by a reaction of 2,4-, 2,5-, 2,6-, or 3,5-pyridine dicarboxylic acid with 3,3',4,4'-tetraaminobiphenyl (TAB) using the PPA Process. Exceedingly pure monomers were required to polymerize the py-PBIs, and IV values of 1.0–2.5 dL g<sup>-1</sup> were obtained. The 2,4- and 2,5-py-PBI membranes formed mechanically strong films, whereas the 2,6-py-PBI membrane was mechanically weak and the 3,5-py-PBI was unable to form films due to high solubility in PPA. All of the py-PBI structures were thermally stable in both nitrogen and air in temperatures up to 420°C. The 2,5- and 2,6-py-PBI were reported as having conductivities of 0.2 S cm<sup>-1</sup> and

0.1 S cm<sup>-1</sup> at 160–200°C, respectively. The 2,5-py-PBI was found to have the most mechanically robust structure. It was hypothesized that the enhancement of mechanical properties was due to its *para*-orientation as opposed to the other py-PBIs having a *meta*-orientation. In addition, the doping level of 2,5-py-PBI averaged 20 mol of phosphoric acid per polymer repeat unit. Because PPA-Processed 2,5-py-PBI was an extremely good candidate for fuel cell testing, polarization tests of the MEA were performed (Fig. 8). The platinum loading on the anode and cathode was 1.0 mg cm<sup>-2</sup> with 30% Pt in Vulcan XC-72 carbon black. The active area for the MEA was 10 cm<sup>2</sup>. The membranes used non-humidified H<sub>2</sub>/O<sub>2</sub> and higher temperatures improved the performances of 2,5-py-PBI MEA.

There have been studies indicating that blends of PBI polymers with pyridine-containing polymers could prove useful in a high-temperature PEM fuel cell. Kallitsis et al. [31] combined commercially supplied *m*-PBI with an aromatic polyether that contained a pyridine moiety in the main chain (PPyPO); these polymer blends were then soaked in 85% wt PA. Dynamic mechanical analysis of a 75/25 PBI/PPyPO



**Polybenzimidazole Fuel Cell Technology. Figure 8**

Polarization curves under hydrogen and oxygen gases at various temperatures of PA-doped 2,5-py-PBI membranes [30]

block copolymer showed reasonable mechanical strength and flexibility. The conductivity of this copolymer was not reported, but the conductivity of 85/15 PBI/PPyPO block copolymer was  $0.013 \text{ S cm}^{-1}$  at a relatively low PA doping level. Further investigation of these systems is required to prove its utility as a fuel cell membrane.

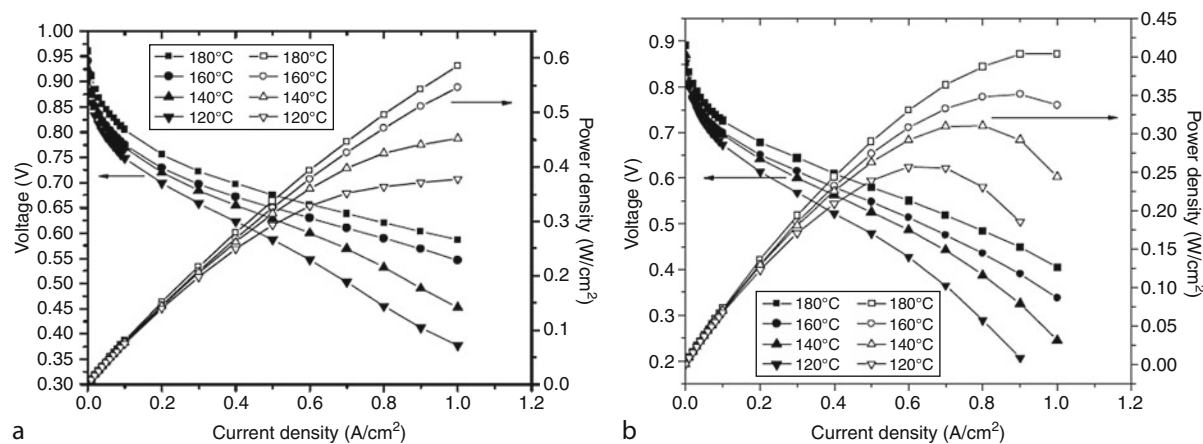
**Sulfonated PBI** Sulfonated aromatic polymers have been widely investigated [32–42] for fuel cell use due to their enhanced physical and chemical robustness, acid and water retention, and conductivity over that of Nafion and other perfluorosulfonic acid-type polymers. Thus, due to the enhanced properties of PBI, it was logical to investigate the physical and chemical properties of sulfonated PBI (s-PBI) membranes. Sulfonation of PBI typically occurs by either direct sulfonation of the polymer backbone [23, 43, 44], grafting sulfonated moieties onto the backbone [23, 45], or by a polycondensation reaction that bonds aromatic tetraamines to sulfonated aromatic diacids [46–48]. Compared to other sulfonation methods, polycondensation reactions provide more control over the degree of sulfonation.

Mader investigated the physical and chemical properties of s-PBI with PA as the dopant (Fig. 5e) [48]. The polymer was synthesized by two different synthetic pathways; the first was a direct polycondensation reaction of 2-sulfoterephthalic acid (s-TPA) and TAB using

the PPA Process, and the second was a post-sulfonation reaction of *p*-PBI using concentrated sulfuric acid. The IV values for the polymer membranes derived from the polycondensation reaction ranged from 1 to  $2 \text{ dL g}^{-1}$ ; these polymers had sufficiently high molecular weights to allow strong films to be cast. In addition, these polymer membranes could achieve doping levels between 28 and 53 mol PA/PRU, which resulted in significantly high conductivity values (all above  $0.1 \text{ S cm}^{-1}$  at all temperatures between 100 and  $200^\circ\text{C}$ ).

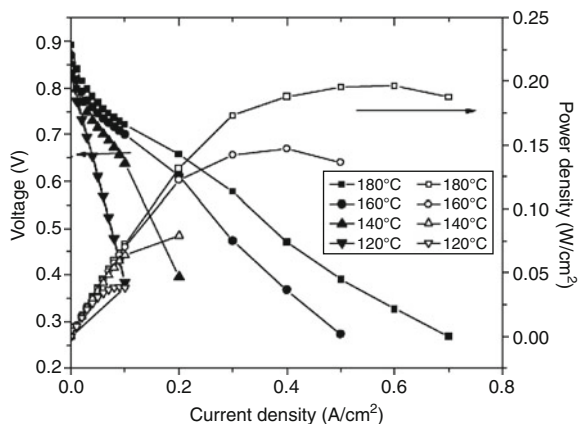
Based on the preliminary data, s-PBI polymer membranes were excellent candidates for fuel cell tests. Polarization tests were run using an s-PBI membrane with an IV value of  $1.71 \text{ dL g}^{-1}$ , a PA doping level of 52.33 mol PA/PRU, and a conductivity of  $0.248 \text{ S cm}^{-1}$ ; the results are depicted in Fig. 9. The s-PBI membrane exhibited its highest performance at  $160^\circ\text{C}$ , producing 0.6788 V at a current density of  $0.2 \text{ A cm}^{-2}$ . This performance compares well to that of other PBIs produced by the PPA Process, which is typically around 0.6–0.7 V at  $0.2 \text{ A cm}^{-2}$ .

The s-PBI homopolymer was shown to have both excellent resistance to gas impurities and excellent longevity. A reformate gas composed of 70% hydrogen, 28% carbon dioxide, and 2% carbon monoxide was used as the fuel while air was used as the oxidant. As depicted in Fig. 10, the fuel cell performance increased with increasing temperature; this is explained by the retardation of carbon monoxide poisoning that occurs



**Polybenzimidazole Fuel Cell Technology. Figure 9**

Polarization curves (filled symbols) and power density curves (unfilled symbols) of s-PBI using (a) hydrogen and oxygen and (b) hydrogen and air [48]



**Polybenzimidazole Fuel Cell Technology. Figure 10** Polarization curves (filled symbols) and power density curves (unfilled symbols) of *s*-PBI using reformat and air [48]

at high temperatures. The performance loss of *s*-PBI MEA was measured by holding the MEA at  $0.2 \text{ A cm}^{-2}$  at  $160^\circ\text{C}$  for 1,200 h using  $\text{H}_2/\text{O}_2$ . After reaching stabilization at the 343rd hour, the MEA had a voltage loss of  $0.024 \text{ mV h}^{-1}$  for the remainder of the test.

Mader also investigated *s*-PBI/*p*-PBI random copolymers (Fig. 5f) for use in fuel cells [48]. The random copolymer was synthesized by reacting TAB, TPA, and *s*-TPA in a reaction flask and the membrane was cast via the PPA Process. High molecular weight polymers were achieved with IV values exceeding  $1.8 \text{ dL g}^{-1}$ ; this allowed for mechanically strong films to be cast. As the ratio of *s*-PBI/*p*-PBI decreased, the molecular weight of the polymer proportionally increased. Higher PA loading was seen at lower *s*-PBI/*p*-PBI ratios, indicating a stronger attractive force between PA and *p*-PBI than PA and *s*-PBI. The PA loading values almost directly corresponded to the conductivity of the membranes. The 75/25 *s*-PBI/*p*-PBI membrane had a PA loading value of  $20.32 \text{ mol PA/PBI}$  and a conductivity of  $0.157 \text{ S cm}^{-1}$ , whereas the 25/75 *s*-PBI/*p*-PBI membrane had a PA loading value of  $40.69 \text{ mol PA/PBI}$  and a conductivity of  $0.291 \text{ S cm}^{-1}$ .

Fuel cell performance tests were conducted on the random copolymers. Even though the 25/75 *s*-PBI/*p*-PBI random copolymer had a higher conductivity than that of *p*-PBI homopolymer, it was found that all of the random copolymers showed lower performance than *p*-PBI. The 50/50 and 75/25 *s*-PBI/*p*-PBI random copolymers had lower performance than the *s*-PBI

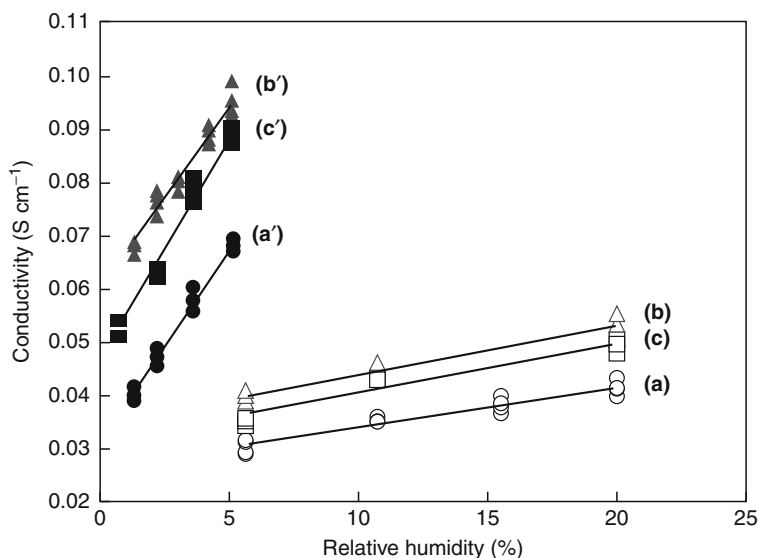
homopolymer at all PA doping levels. However, the 25/75 *s*-PBI/*p*-PBI random copolymer performed comparably to the *s*-PBI homopolymer at equivalent PA doping levels.

**PBI-Inorganic Composites** For conventionally prepared PBI membranes, as the acid doping levels of PBIs increase, the conductivity and overall performance of the PBI membranes also tend to increase. However, as high acid doping levels are reached for PBI membranes, the mechanical strength of the membrane significantly decreases. Inorganic fillers for PBI membranes have been investigated to improve membrane film strength, thermal stability, water and acid uptake, and conductivity. These composite membranes have only been examined using *m*-PBI and the conventional casting method.

He et al. investigated the use of zirconium phosphate (ZrP) in a PA/PBI system [49]. The conductivity of *m*-PBI with a doping level of 5.6 PA/PRU increased from  $0.068 \text{ S cm}^{-1}$  to  $0.096 \text{ S cm}^{-1}$  with the addition of 15 wt% ZrP at  $200^\circ\text{C}$  and at 5% relative humidity. As seen in Fig. 11, the conductivity of the membrane increased as the relative humidity and temperature of its environment increased. Conductivities of other inorganic fillers, such as phosphotungstic acid, silicotungstic acid, and tricarboxylbutylphosphonate, are comparable or lower than that of ZrP. Unfortunately, there have been no fuel cell performance tests published on these systems. Overall, these inorganic fillers improved the conductivity of *m*-PBI membranes.

**Other Modified PBIs** Multitudes of other organically modified PBI membranes exist that include, but are not limited to, fluorinated PBI, ionically and covalently cross-linked PBI, PBI blends, and a wide variety of PBI copolymers. Because there are far too many to describe, this subsection will highlight select PBI membranes that have not been included in the prior subsections.

Qian et al. investigated the use of hexafluoroisopropylidene-containing polybenzimidazole (6F-PBI, Fig. 5g) in fuel cells [50]. The polymer was synthesized via the PPA Process through the reaction of TAB with 2,2-Bis(4-carboxyphenyl) hexafluoropropane in PPA. High molecular weight polymer with an IV value of  $0.98 \text{ dL g}^{-1}$  was achieved. Although the PA doping level



**Polybenzimidazole Fuel Cell Technology. Figure 11**

Conductivity study of ZrP/*m*-PBI system for (a) *m*-PBI at 140°C, (a') *m*-PBI at 200°C, (b) 15 wt% ZrP in *m*-PBI at 140°C, (b') 15 wt% ZrP in *m*-PBI at 200°C, (c) 20 wt% ZrP in *m*-PBI at 140°C, and (c') 20 wt% ZrP in *m*-PBI at 200°C [49]

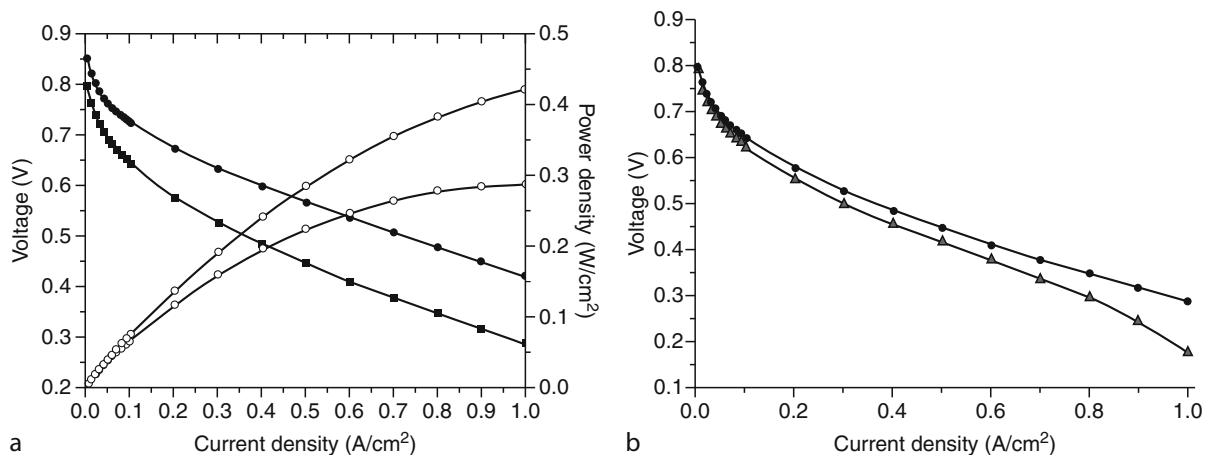
of 6F-PBI was considerably high (30–40 mol PA/PRU), the membrane only achieved a peak conductivity value of  $0.09 \text{ S cm}^{-1}$  at  $180^\circ\text{C}$ . This is lower than that of PPA-Processed *p*-PBI that achieved approximately  $0.25 \text{ S cm}^{-1}$  at  $160^\circ\text{C}$ .

The mechanical strength of 6F-PBI at high PA doping levels was strong enough to fabricate a membrane for fuel cell testing. Polarization and power density curves of 6F-PBI using hydrogen and reformat gases as fuel are illustrated in Fig. 12. Using hydrogen as fuel and air as the oxidant, the 6F-PBI MEA achieved a steady-state voltage of 0.58 V at a current density of  $0.2 \text{ A cm}^{-2}$ . When oxygen was used as the oxidant at the same current density, the steady-state voltage increased to 0.67 V. Additionally, the MEA showed excellent resistance to carbon monoxide poisoning. When a reformat gas comprised of 40% hydrogen, 40.8% nitrogen, 19%  $\text{CO}_2$ , and 0.2% CO was used as fuel and air was used as the oxidant, the CO poisoning effects produced an approximate 3 mV reduction in voltage. This study illustrates that low levels of CO poisoning have little effect on the 6F-PBI MEA operating at this temperature.

Commonly known as 2OH-PBI (Fig. 5h), poly(2,2'-(dihydroxy-1,4-phenylene)5,5'-bibenzimidazole) is another PBI membrane with extremely promising

properties. Yu and Benicewicz [51] synthesized 2OH-PBI homopolymer by combining TAB with 2,5-dihydroxyterephthalic acid (2OH-TPA) in PPA and cast it via the PPA Process. Yu also synthesized the 2OH-PBI/*p*-PBI random copolymer by reacting both 2OH-TPA and TPA simultaneously with TAB; the copolymer membrane was also cast using the PPA Process. It was proposed that the 2OH-PBI homopolymer was significantly cross-linked through phosphoric acid ester bridges. Because of this cross-linking, the polymer was unable to be dissolved and an IV value could not be determined. Upon hydrolysis of the ester bridges by sodium hydroxide, the IV value of the homopolymer was measured as  $0.74 \text{ dL g}^{-1}$ . The acid doping level of 2OH-PBI homopolymer was approximately 25 PA/PRU, and its conductivity at  $160^\circ\text{C}$  was  $0.35 \text{ S cm}^{-1}$ . It is important to note that at all temperatures between room temperature and  $180^\circ\text{C}$ , the conductivity of 2OH-PBI homopolymer was greater than that of *p*-PBI. As the ratio of 2OH-PBI/*p*-PBI decreased in the random copolymer, the doping level and conductivity decreased. It was found that the conductivity of the material was highly dependent on the chemical structure of the PBI membrane and not just the doping level.

Using a Pt anode electrode and a Pt-alloy cathode electrode, polarization tests were performed



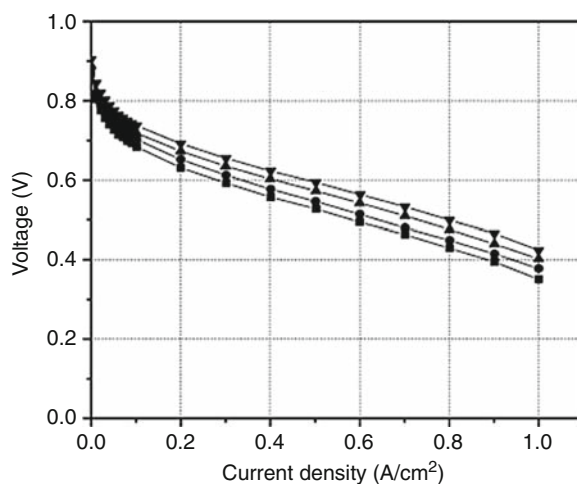
**Polybenzimidazole Fuel Cell Technology. Figure 12**

Graph (a) Polarization curves (filled symbols) and power density curves (unfilled symbols) of 6F-PBI using H<sub>2</sub>/air (squares) and H<sub>2</sub>/O<sub>2</sub> (circles). Graph (b) Polarization curves of 6F-PBI using H<sub>2</sub>/air (circles) and reformat/air (triangles) [50]

on the homopolymer 2OH-PBI MEA (Fig. 13). The homopolymer produced a voltage of 0.69 V using a load of 0.2 A cm<sup>-2</sup> at 180°C and H<sub>2</sub>/air; this is greater than the 0.663 V produced by *p*-PBI under the same conditions. The high acid doping level and the membrane chemistry significantly contribute to the excellent performance of the 2OH-PBI membrane. Overall, the fuel cell performance of 2OH-PBI is comparable to that of *p*-PBI.

Segmented PBI block copolymers have also been explored for fuel cell use [52]. Scanlon synthesized a 52/48 *p*-PBI/*m*-SPBI (Fig. 5i) segmented block copolymer by polymerizing the oligomer of *p*-PBI with that of *m*-SPBI. The oligomers were polymerized in PPA and cast by the PPA Process. Even with an extremely high PA doping level of 91.5 mol PA/PRU, the polymer film had very strong mechanical properties. Under low humidity at 160°C, the segmented copolymer achieved a conductivity of 0.46 S cm<sup>-1</sup>. Because of the great results, a *p*-PBI/*m*-SPBI MEA was constructed for use in fuel cell performance tests. The polarization curves of the segmented copolymer MEA displayed a voltage of 0.62 V at 0.2 A cm<sup>-2</sup> at 160°C and 0.65 V at 0.2 A cm<sup>-2</sup> at 200°C. As implied by the data, these membranes are excellent candidates for high-temperature fuel cells.

**Membrane Electrode Assembly Durability** As explained in a previous section, a membrane electrode



**Polybenzimidazole Fuel Cell Technology. Figure 13**

Polarization curves of 2OH-PBI using hydrogen as the fuel and air as the oxidant at 120°C (squares), 140°C (circles), 160°C (triangles), and 180°C (down-triangles) [51]

assembly (MEA) consists of the polymer membrane that is sandwiched between an anode and a cathode electrode, respectively. The electrodes are composed of a conductive carbon network that supports a catalyst on a gas diffusion layer. An additive, such as polytetrafluoroethylene (PTFE), helps bind the Pt/C catalyst to the gas diffusion layer. At the anode, the catalyst facilitates the oxidation of hydrogen into its constituent electrons and protons. As the protons are passed



through the acid-doped membrane to the cathode, the electrons are passed through an external circuit, thereby creating electricity. Finally, the electrons and protons react with oxygen at the cathode electrode to form water as the final reaction product.

Although PBI membranes are highly resistant to degradation, it is possible for the membranes to fail. Common degradation modes for PBI membranes at operating temperatures of 120–200°C include membrane thinning and pin-hole formation. If there is too much pressure on the membrane, phosphoric acid could be pushed out of the polymer matrix and “thin out” the membrane. An extreme occurrence of membrane thinning results in pin-hole formations. Both of these occurrences result in increased fuel crossover and reduced fuel cell efficiency. Firm gasket materials help to evenly distribute pressure and prevent over-compression of the membrane [53].

The catalyst-coated electrodes of the MEA must be extremely durable in the presence of harsh physical and chemical environments. The oxidation and reduction processes create immense stress on the electrodes and trigger physical and chemical reactions to occur. A summary of the main MEA and component degradation modes have been previously reported [53, 54]. By means of electrochemical Ostwald ripening, Pt-metal agglomeration causes the loss of electrochemical surface area and decrease of reaction kinetics mainly through a dissolution-recrystallization process [55–57]. Oxidation reactions can also cause corrosion of the gas diffusion layer and carbon components in the electrodes, which would result in acid flooding, an increase in mass transport overpotentials, a decrease reaction kinetics and also, most severely, the loss of the mechanical integrity of the electrodes. Phosphoric acid can dissolve the Pt-metal catalyst and phosphoric acid anions ( $\text{H}_2\text{PO}_4^-$ ) could adsorb onto the catalyst surface; both of these events would decrease the electrochemical surface area and reaction kinetics. In addition, phosphoric acid evaporation from the catalyst layer would result in similar consequences.

Typical commercial gas diffusion electrodes contain high-surface area carbon supported catalysts, e.g., Pt/Vulcan XC 72. Platinum is typically used as the catalyst at both the anode and cathode electrodes because it facilitates the reduction and oxidation reactions at high efficiency. However, due to the degradation modes

previously mentioned, performance of the catalyst is lost over time. Novel platinum-based catalysts have been developed to increase the stability of the electrode catalysts. Compared to a commercial Pt/C (46.6 wt% TKK),  $\text{Pt}_4\text{ZrO}_2/\text{C}$  catalysts have been shown to decrease the overall performance loss of the MEA [58]. The  $\text{Pt}_4\text{ZrO}_2/\text{C}$  catalyst showed a higher resistance to Pt-sintering than Pt/C following 3,000 cycles of a potential sweep test between 0.6 and 1.2 V versus reversible hydrogen electrode ( $20 \text{ mV s}^{-1}$ ). The  $\text{ZrO}_2$  is thought to act as an anchor to slow the agglomeration of platinum particles.

In order to improve especially the cathode catalyst kinetics and the catalyst stability, alloying of Pt with a base metal such as nickel or cobalt is widely done. Origins of these alloys date back to early phosphoric acid fuel cell development [59]. These alloys have been reported to typically improve the cathode kinetics for oxygen reduction by roughly 25–40 mV [59] or a factor of 1.5–4 when considered reaction rates. Commercial MEAs using PBI-based membranes also use Pt-base metal alloy catalyst on the cathode [60, 61]. The origin of the kinetic improvements for the Pt-base metal alloys is discussed manifold in literature [62–70]: (1) modification of the electronic structure of Pt (5-d orbital vacancies), (2) change in the physical structure of Pt (Pt–Pt bond distance and coordination number), (3) adsorption of oxygen-containing species from the electrolyte onto the Pt or alloying element, and/or (4) redox-type processes involving the first-row transition alloying elements. However, as discussed in detail in the recent work by Stamenkovic et al. [70], the main effect is a shift of the Pt d-band center to lower energy values which induces a surface which adsorbs oxygenated and spectator species to a lower extent and therefore makes more active sites available for the oxygen reduction to proceed.

Other additives to platinum-based electrodes, such as tin-oxide ( $\text{SnOx}$ ) [71], have also been shown to significantly improve the catalytic activity of the oxygen reduction reaction. Using a PPA-Processed *m*-PBI membrane with a 7 wt% SnO in  $\text{Pt/SnO}_2/\text{C}$  catalyst under unhumidified  $\text{H}_2/\text{O}_2$  at 180°C, a voltage of 0.58 V under a load of  $0.2 \text{ A cm}^{-2}$  was produced. Under the same conditions, a *m*-PBI MEA using a Pt/C catalyst produced only 0.4 V at  $0.2 \text{ A cm}^{-2}$ .

PBI has also been investigated as an additive to platinum-based electrodes. It is thought that incorporation of PBI in the catalyst layer would provide a better interface for proton conduction between the electrode and membrane. Qian [72] incorporated 6F-PBI into the electrodes by four different methods: formation of a PBI bilayer inserting a thin 6F-PBI membrane between an E-TEK cathode and *p*-PBI membrane, casting 6F-PBI/PPA directly onto the E-TEK electrodes and hydrolyzing to form the gel, spraying a 6F-PBI/DMAC solution onto the electrodes, and coating the electrodes with a mixture of 6F-PBI and catalyst (the PBI replaced PTFE). The bilayer method decreased fuel cell performance, and it is proposed that this occurred by creating a large interface resistance between the two PBI membranes. Both the casting method and the spraying method improved electrode kinetics, and it is postulated that this occurred due to a lower interface resistance. In addition, a significant decrease in fuel cell performance showed that 6F-PBI could not be used to replace PTFE.

As an outlook to further improvements of catalyst kinetics and durability in low- and high-temperature polymer electrolyte fuel cells, several possibilities are currently under investigation [73]: (1) extended large-scale Pt and Pt-alloy surfaces [70]; (2) extended nanostructured Pt and Pt-alloy films [74]; (3) de-alloyed Pt-alloy nanoparticles [75]; (4) precious metal free catalyst as described by Lefevre et al. [76], e.g., Fe/N/C catalysts; and (5) additives to the electrolyte which modify both adsorption properties of anions and spectator species and also the solubility of oxygen [77]. The latter approach is specific to fuel cells using phosphoric acid as electrolyte.

### PBI-PA Fuel Cell Systems and Their Applications

*Para*-PBI is one of the most common polymers used in commercial PBI-based fuel cell systems. A mechanically strong and chemically stable polymer, *p*-PBI has proved to be one of the most reliable PBI polymers for MEA use. Load, thermal, and shutdown–startup cycling tests performed on the *p*-PBI MEA indicated that high temperatures (180°C and 190°C) and high load conditions slightly increased PA leaching from the MEA system. However, at steady-state fuel cell operation at 80–160°C studies showed that PA loss would not

be a significant factor in fuel cell degradation [54, 78]. Long-term studies showed minimal performance degradation over a 2-year span and indicated excellent commercial fuel cell potential [53]. Compared to state-of-the-art phosphoric acid PEMFCs [79], evaporation of phosphoric acid from commercial PBI-based Celtec P1000 MEAs is reduced by a factor of roughly 2–3. This is a key factor of long-term stable operation for PBI-based fuel cells.

For the transition of PBI-based fuel cell science into commercial products, the appropriate manufacturing processes need to be developed. Most companies rely on manual operations [80] for PBI-based MEA fabrication. Only recently have significant efforts been devoted to developing automated production lines because simple changes in MEA materials and architecture could necessitate the use of different manufacturing equipment. To accommodate the evolution of fuel cell science, a flexible modular manufacturing line has been developed. Since 2002, BASF Fuel Cell GmbH (previously PEMEAS) has used the line to accommodate three generations of MEAs. The details of this manufacturing process will be further discussed in section “[Advances in PBI MEA Manufacturing](#)”.

Commercial PBI-based high-temperature PEMFCs provide energy to a wide array of electronic devices. Hydrogen fuel cell vehicles, both for the private consumer and public transportation, are growing in popularity as pollution and fossil fuel prices continue to increase. Hydrogen offers 2–3 times the overall efficiency in a fuel cell as gasoline does in a typical combustion engine [81]. High-temperature fuel cells are also popular as backup generators and combined heat and power devices for stationary use. These types of systems typically produce 1–10 kW, which is enough energy to power a house or a multifamily dwelling. In addition to providing energy, combined heat and power devices use waste heat to heat water and preheat the fuel cell system components, thereby increasing the overall efficiency of the fuel cell system. Fuel cells also offer applications in mobile electronic devices such as laptops and cell phones. Commonly coupled with a methanol reformer, these fuel cell systems are remarkably portable and can power electronics for hours of continuous use.

In addition to producing electricity, these PEMs have been used as a purification device for hydrogen gas.

Consider the purification device to have the same basic architecture as a fuel cell. A platinum catalyst splits contaminated hydrogen gas into protons and electrons at the anode. Using an external power source, the electrons are driven through an external circuit to the cathode while the protons are allowed to transport across the membrane from the anode to the cathode. The electrons and protons recombine, thereby creating a higher grade hydrogen gas at the cathode while leaving behind the undesired constituents at the anode. These hydrogen pump devices will be further discussed in section “Hydrogen Pump”.

### In-Depth Analysis of PPA-Processed *p*-PBI MEA

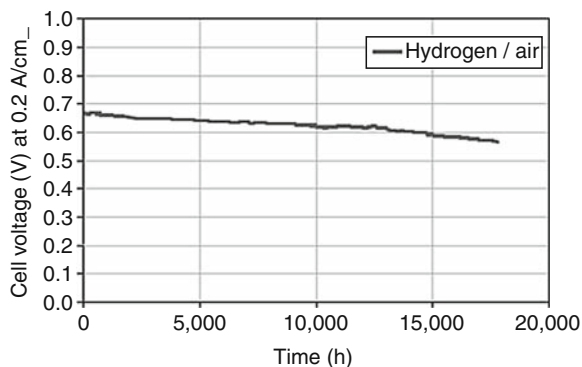
PBI-based high-temperature MEAs offer many benefits over more well-known perfluorosulfonic acid Nafion PEMs. Unlike low-temperature Nafion MEAs, high-temperature PA-doped PBI membranes do not need to be hydrated, and therefore, do not require an external humidification of the gases. Additionally, running at high temperatures generally improve electrode kinetics and proton conductivities while requiring smaller heat exchangers. For PBI fuel cell science to transition into commercially available products, the reliability of PBI fuel cell stacks needs to meet specific requirements. The Department of Energy (DOE) specified durability targets of >5,000 h (>150,000 miles) of automotive fuel cell operation and >40,000 h for stationary applications for 2010. Primarily, the durability of the fuel cell stack dictates the durability of the entire system [82]. In-depth durability studies of PBI MEAs have been performed [53, 54, 61, 78, 83–85] to evaluate the viability of commercial PBI fuel cells. In addition to fuel impurity and PA retention tests, load, thermal, and shutdown–startup cycling tests are commonly performed to evaluate the MEAs.

*p*-PBI MEAs have displayed a relatively high resistance to carbon monoxide and sulfur contaminants [78, 83, 86, 87]. While Nafion and other traditional low-temperature PEM fuel cells are often poisoned by small amounts of carbon monoxide (5–50 ppm) in the fuel or oxidant, *p*-PBI and other PBI membranes have been shown to perform with minimal voltage loss in  $10^4$  ppm of carbon monoxide. Operating the fuel cell at  $180^\circ\text{C}$  with a load of  $0.2\text{ A cm}^{-2}$  with a reformat gas (70%  $\text{H}_2$ , 1.0%  $\text{CO}$ , and 29%  $\text{CO}_2$ ), the voltage loss was

only 24 mV as compared to pure hydrogen [27]. This decrease in voltage occurred as a result of fuel dilution and carbon monoxide poisoning. As explained in section “Introduction to Polybenzimidazole Fuel Cell Sustainability”, the cell is able to resist poisoning because the high operating temperatures allow for reversible binding of carbon monoxide from the catalyst. Details on the CO adsorption isotherms in the presence of hydrogen under fuel cell operation conditions between  $150^\circ\text{C}$  and  $190^\circ\text{C}$  can be found in literature [87]. Similarly, Garseny et al. [86] reported that a PBI MEA from BASF Fuel Cell GmbH (Celtec-P Series 1000) is 70 times more resistant to sulfur contaminants than Nafion MEAs. Using air contaminated with 1 ppm  $\text{H}_2\text{S}$  or  $\text{SO}_2$  as the oxidant, the performance of Nafion decreased by 82.9% while the performance of the Celtec-P MEA decreased by <2%. Garseny et al. proposed that  $\text{H}_2\text{S}$  is converted to  $\text{SO}_2$ , and that  $\text{SO}_2$  adsorbs onto the Pt catalyst surface. At temperatures above  $140^\circ\text{C}$ , this  $\text{SO}_2$  is desorbed and flushed out of the system. Schmidt and Baurmeister showed that the  $\text{H}_2\text{S}$  tolerance of PBI-based Celtec P1000 MEAs is in the range of 10 ppm [83], a value significantly larger than typical fuel processing catalyst can tolerate. More than 3,000 h operation in reformat with 5 ppm  $\text{H}_2\text{S}$  and 2%  $\text{CO}$  is proven. Overall, *p*-PBI-based fuel cells can resist contaminant poisoning far better than traditional low-temperature PEM fuel cells, an effect which can mainly be ascribed to the operation temperature between  $150^\circ\text{C}$  and  $190^\circ\text{C}$ .

Under continuous operation and appropriate stack design and components, the PBI membranes retain phosphoric acid extremely well. Long-term performance tests show that *p*-PBI fuel cells can operate for over 2 years with minimal performance degradation (Fig. 14). This durability is attributed to the unique nature of PBI membrane formed by the PPA Process, which allows it to retain PA under continuous operating conditions. The amount of PA lost from the *p*-PBI MEA per hour was approximately  $10\text{ ng h}^{-1}\text{ cm}^{-2}$ , which is equivalent to a  $50\text{ cm}^2$  cell losing 8.74 mg PA after 2 years of operation. Such a small loss strongly suggests that the life span of a *p*-PBI PEM fuel cell would not be significantly influenced by PA depletion.

Phosphoric acid loss was also monitored during a selection of dynamic durability tests, including load and thermal cycling tests [78]. A single load cycle test



**Polybenzimidazole Fuel Cell Technology.** Figure 14 Long-term durability test of *p*-PBI MEA at 160°C using hydrogen/air without humidification

involved measuring the voltage at 160°C at three different loads: open circuit voltage (OCV), 0.2 A cm<sup>-2</sup>, and 0.6 A cm<sup>-2</sup>. Air and pure hydrogen were supplied to the MEA as oxidant and fuel, respectively. The voltage of the MEA was measured at OCV for 2 min, followed by 0.2 A cm<sup>-2</sup> for 30 min, and then 0.6 A cm<sup>-2</sup> for 30 min. A total of 500 load cycles were performed on a *p*-PBI MEA, and the results indicated that larger loads corresponded to an increased PA loss rate (approximately 20 ng h<sup>-1</sup> cm<sup>-2</sup>). Thermal cycling tests were performed by measuring the voltage of the MEA with a constant applied current density of 0.2 A cm<sup>-2</sup> while either cycling the temperature between 120°C and 180°C (for a high-temperature cycle) or between 80°C and 120°C (for a low-temperature cycle). Both the high- and low-temperature cycles were performed 100 times each. The results showed that higher temperatures were associated with an increased PA loss rate (almost 70 ng h<sup>-1</sup> cm<sup>-2</sup> for the high-temperature cycle and 20 ng h<sup>-1</sup> cm<sup>-2</sup> for the low-temperature cycle). It was proposed that at the higher load and temperature conditions, more water is generated at the cathode. By means of a steam distillation mechanism, an increased amount of PA is lost from the MEA. As indicated by both cycling tests, phosphoric acid loss becomes a significant factor of cell degradation only under extreme conditions.

Shutdown–startup cycling tests have been extensively studied by Schmidt and Baurmeister of BASF Fuel Cell GmbH [54, 61]. Two PBI-based PEFC Celtec-P1000 MEAs were tested under different

operation modes; one was run under shutdown–startup cycling parameters (12 h shutdown followed by operation for 12 h at 160°C under a load of 0.2 A cm<sup>-2</sup>) while the other was continuously operated at 160°C under a load of 0.2 A cm<sup>-2</sup>. Both MEAs were operated for more than 6,000 h, during which the shutdown–startup cycling MEA underwent more than 270 cycles. While the continuously operating MEA had an average voltage degradation rate of roughly 5 μV h<sup>-1</sup>, the cycling MEA averaged a voltage degradation of 11 μV h<sup>-1</sup> or 0.2 mV cycle<sup>-1</sup>. This increase in voltage degradation was attributed to an increased corrosion of the cathode catalyst support, thereby significantly increasing the cathodic mass transport overpotential. The observed corrosion was a result of a reverse-current mechanism that occurs under shutdown–startup cycling conditions [88].

Illustrated by the previously discussed durability tests, *p*-PBI MEAs have been shown to be physically and chemically robust. Highly resistant to fuel contaminants, PBI MEAs are resistant to poisoning effects that would typically expunge a low-temperature Nafion fuel cell system. Long-term steady-state and dynamic durability tests showed that PA loss typically is not a cause of cell degradation. Additionally, Schmidt and Baurmeister showed that PBI MEAs are susceptible to cell degradation under extreme shutdown–startup conditions. Overall, *p*-PBI MEAs have exhibited much potential for use in fuel cell systems.

### Advances in PBI MEA Manufacturing

As previously discussed, the manufacturing processes of PBI-based fuel cells need to be improved to make fuel cells a viable commercial product. To put this requirement into perspective, the US Department of Energy has set a goal of producing 500,000 fuel cell cars each year. If these vehicles are powered using current *p*-PBI membranes, this goal requires the production of seven MEAs per second and approximately 250,000 m<sup>2</sup> of electrode per day. Additionally, the performance of each of these MEAs would need to be tested; this is a process called “burn-in testing.” A typical test stand is 25 ft<sup>2</sup>, costs roughly \$50,000, and can only test one stack of MEAs at a time. If each stack requires a 24 h burn-in test, the test facility size would exceed 34,000 ft<sup>2</sup> and house equipment would cost over

\$68.5 million. Existing manufacturing processes need to be improved in order to reach this goal.

The Center for Automation Technologies and Systems (CATS) at Rensselaer Polytechnic Institute has developed a flexible manufacturing process for BASF Fuel Cell GmbH to accommodate the evolving science of fuel cells [89–91]. If one changes the fuel cell type, size, materials, MEA architecture, design, or application, the manufacturing line could be significantly affected. Therefore, a modular manufacturing line was developed by CATS in 2002 that could produce a large range of MEA sizes (1–1,000 cm<sup>2</sup>), could handle a wide variety of materials (membranes, gaskets, electrodes, etc.), could assemble these materials in different architectures, and could be expanded to integrate additional systems. Each module could be singularly operated or could operate as a subset of the entire process; this modular construction is shown in Fig. 15. Over the past 8 years, this manufacturing line has evolved over three generations of MEA devices.

Members of CATS continue to make great strides in order to reduce costs and improve the overall efficiency of MEA fabrication. Laser cutting and joining of the PBI membranes both uses less power and delivers

tighter tolerances than that of conventional cutting and joining. Ultrasonic technology has also been explored to replace the thermal joining of the three components of an MEA. Preliminary results exhibited a significant reduction in pressing time by approximately 90% in addition to using less energy. Additionally, an automated visual inspection of the MEA has been developed using a high precision motion system, multiple cameras and lighting equipment, and software MAT-LAB 7.0 with Image Processing Toolbox [89]. As fuel cell science continues to evolve, so will the manufacturing processes.

### Combined Heat and Power

Stationary combined heat and power (CHP) devices are often considered the primary application of high-temperature PBI-based fuel cells. These devices are used to provide both electricity and heat (in the form of hot air or water) to small-scale residential homes or large-scale industrial plants using hydrogen derived from the widely distributed natural gas network. PBI MEAs are ideally situated for combined heat and power devices because they efficiently provide electricity while generating heat as a by-product. Furthermore, these



**Polybenzimidazole Fuel Cell Technology. Figure 15**

A portion of the 2002 pilot line depicting its modular construction [90]

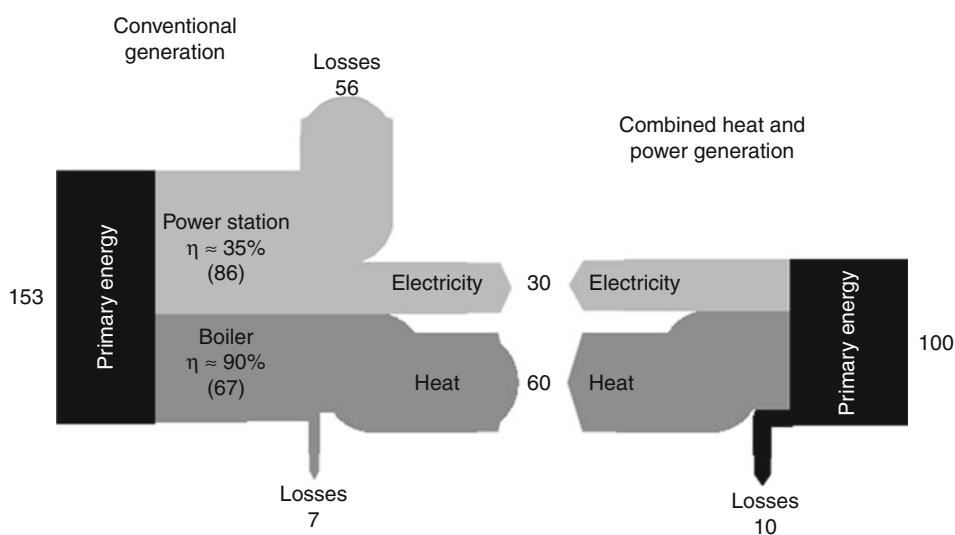
devices could be used to provide reliable backup power to residential homes, hospitals, servers, etc.

J.-Fr. Hake et al. [92] compared the conventional generation of heat and electricity to that of small-scale combined heat and power generation by high-temperature fuel cells, and the results of which are shown in Fig. 16. The small-scale CHP devices studied were used to provide electricity, space heat, and warm water to both residential and commercial buildings. The conventional generation of electricity is much less efficient than that of small-scale CHP devices due to the issues of transportation and storage. In addition to efficiently converting chemical energy into electrical energy, CHP fuel cell systems further act as a sustainable energy conversion device by reducing the total amount of greenhouse gas emissions. Hake et al. considered the penetration of small-scale CHP fuel cell technology into the US residential sector market starting in 2014 until a saturation point as a logarithmic function. To improve the accuracy of the study, Hake considered the trends of the Japanese small-scale CHP market [92, 93]. A typical CHP device in Japan costs roughly \$30,000, but analysts expect the price to drop to \$5,000 within 5 years. Analysts also claim that by the year 2050, one in four homes in Japan will run on fuel cells. Also considering current CO<sub>2</sub>

emissions, Hake et al. concluded that adoption of this technology in the USA could reduce CO<sub>2</sub> emissions by up to approximately 50 million tons by 2050; this corresponds to a 4% reduction in the residential sector.

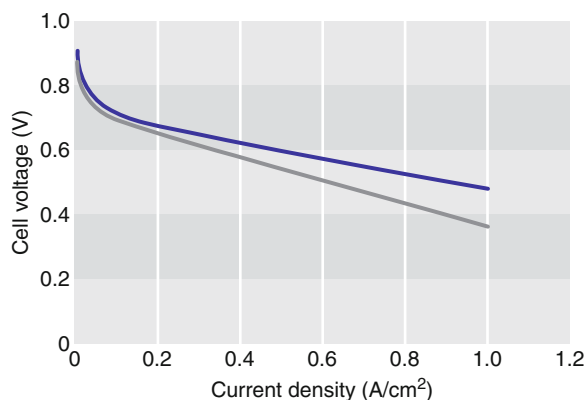
As the largest producer of PBI MEAs, BASF Fuel Cell (previously PEMEAS) produces *p*-PBI PEM MEAs for a wide variety of fuel cell applications. The Celtec<sup>®</sup>-P1000 PEM MEA is typically integrated into either backup or auxiliary power units and can produce from 0.25 to 10 kW. The MEA is also advertised as maintaining performance for over 20,000 h with only a 6 μV h<sup>-1</sup> voltage drop at 160°C [94]. The Celtec<sup>®</sup>-P2100 PEM MEA is used in stationary CHP systems and is capable of producing 0.74–10 kW. The MEA has a long-term stability of over 20,000 h under both steady-state and cycling conditions (300 shutdown–startup cycles with 13 μV h<sup>-1</sup> voltage drop). Polarization curves of a Celtec<sup>®</sup>-P MEA at 160°C using an active area of 45 cm<sup>2</sup> are shown in Fig. 17. PBI-based CHP devices are commercially available from a variety of companies, including Serenergy, Plug Power, and ClearEdge Power. All of these companies assemble a variety of fuel cell devices using commercially available PBI MEAs.

Plug Power of Latham, New York produces a line of PBI-based small-scale CHP devices including the



**Polybenzimidazole Fuel Cell Technology. Figure 16**

Side-by-side comparison of conventional generation of heat and electricity to fuel cell combined heat and electricity generation [92]



**Polybenzimidazole Fuel Cell Technology. Figure 17** Polarization curves of a Celtec<sup>®</sup>-P MEA [94]. The blue line represents using hydrogen/air as fuel/oxidant. The gray line represents a steam reformat of 70% H<sub>2</sub>, 29% CO<sub>2</sub>, and 1% CO/air

GenSys Blue (Fig. 18) [95]. The GenSys Blue is capable of producing 0.5–5 kW of continuous output and is capable of reducing home energy costs by 20–40%. An autothermal (ATR) reformer reacts natural gas (methane) with oxygen and carbon dioxide to produce hydrogen gas that fuels the PEM stack. An inverter is used to improve the efficiency of the CHP device by specifically supplying enough energy to power the home, thereby minimizing energy losses and reducing CO<sub>2</sub> emissions by 25–35%. Additionally, an integrated peak heater ensures proper heating of the entire home.

Serenergy, which is based in Hobro, Denmark, also produces PBI-based fuel cell CHP devices [96]. Serenergy's Serenus 166 Air C v2.5 and 390 Air C v2.5 micro-CHP modules nominally produce 1 and 3.5 kW, respectively. While the 166 model is comprised of one MEA stack of 65 cells, the 390 model uses three MEA stacks each with 89 cells. Both of these systems are able to tolerate fuel impurities up to 5% CO concentrations and 10 ppm H<sub>2</sub>S at 160°C. Because the excess energy can be used to heat up air or water, Serenergy claims that over 80% of the total heat and power generated can be used and that the system efficiency is as high as 57% (the efficiency data was not available). These systems can also be used as auxiliary energy conversion devices.

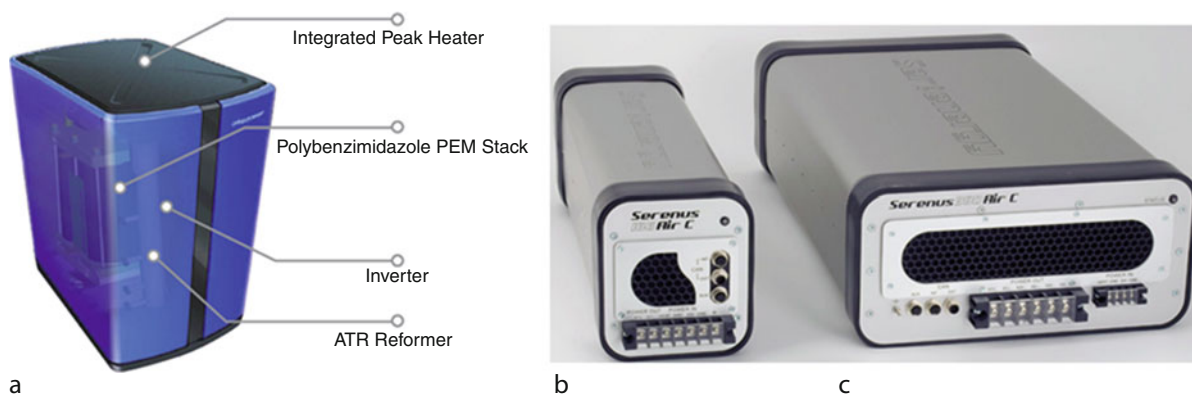
ClearEdge Power also produces a line of small-scale CHP devices, one of which is the ClearEdge5 [97]. Capable of producing 5 kW h<sup>-1</sup> and up to 20,000

BTU h<sup>-1</sup> while running at 150°C, the ClearEdge5 couples a methane reformer to a PBI fuel cell stack using MEAs provided by BASF Fuel Cell GmbH. ClearEdge advertizes that the CHP device can reduce utility bills by up to 50% and cut CO<sub>2</sub> emissions by over 33%. Annually, the device is capable of producing 43,000 kWh in electricity and 50,000 kWh (equivalent) in heat. Similar to other CHP devices, the ClearEdge5 offers at-home production of energy, thereby eliminating the losses associated with transferring the energy. The device is monitored in real-time by ClearEdge Power and can be monitored directly from the owner's smartphone.

### Automotive Transportation

Producing 1.9253 billion metric tons of carbon dioxide, which is roughly 33% of USA's total carbon dioxide emissions, the transportation sector was the largest contributor to pollution in 2008 [98]. According to another 2008 study by the US Department of Energy [2], all transportation in the USA consumed approximately 27.8 quadrillion BTUs. Considering both of these facts, one can conclude that a more sustainable energy source could significantly reduce the carbon footprint of the transportation sector. For the transition of fuel cell science into a viable commercial product to occur, the US Department of Energy has set numerous targets for automotive fuel cell systems. Because a typical internal combustion engine costs roughly \$25–35/kW, a fuel cell system will need to cost roughly \$30/kW to become competitive enough to penetrate the US market. Furthermore, the system must be durable enough to operate for at least 5,000 h (or roughly 150,000 miles). Additional issues of system size and management of air, heat, and water will also play a role in automotive fuel cell viability.

Over the past decade, fuel cell technology has been adapted by the major automotive industries as a cleaner, more efficient method of providing energy to vehicles. In addition to the issue of fuel cell automotive viability, issues of hydrogen sources, hydrogen storage, and fueling stations continue to be addressed and solved. The California Fuel Cell Partnership (CaFCP) is a collaboration of auto manufacturers, energy providers, government agencies, and fuel cell technology companies to promote the commercialization of fuel cell vehicles [81]. In 2009, California had



**Polybenzimidazole Fuel Cell Technology. Figure 18**

Plug Power's GenSys blue (a), Serenergy's Serenus 166 air C v2.5 (b), and Serenergy's Serenus 390 air C v2.5 (c) CHP fuel cell devices [95, 96]

only six public hydrogen fueling stations that were used to fuel roughly 200 vehicles. The CaFCP predicts that in 2014, approximately 5,800 kg of hydrogen will be used to fuel roughly 4,310 fuel cell passenger vehicles and 60 fuel cell busses daily. To accommodate the needs of fuel cell vehicle operators, CaFCP proposed the establishment of 46 new fueling stations by the year 2012. Considering each new station would cost in the range of \$1.5–\$5.5 million, a predicted \$180 million would have to be spent on the fueling station project. Although this “Action Plan” did not specify an amount, this project will provide many new jobs to US residents. The hydrogen used to fuel these stations can be domestically produced as either a low-carbon fuel or potentially as a zero-carbon fuel when produced from renewable sources (such as splitting water into oxygen and hydrogen with solar energy). According to California regulations, at least 33% of the hydrogen must come from such renewable sources.

SunHydro, one of the world's first hydrogen fueling station chains, has set a goal of providing fueling stations along the entire east coast of the USA [99]. Using solar cell technology, every SunHydro station will harvest solar energy to electrolytically split water into hydrogen and oxygen gases. This process is extremely sustainable and will create much less greenhouse gas emissions. This hydrogen highway will stretch from Scarborough, ME, to Miami, FL, and consist of 11 stations. Each station will cost an estimated \$2–\$3 million to construct and will be paid for by private funders.

Over the past decade, many automotive and fuel cell industries have used PBI technology in the development of fuel cell vehicles. In November of 2008, Volkswagen unveiled the VW Passat Lingyu at a Los Angeles Auto Show [100]. The VW Lingyu uses an AB-PBI-based fuel cell stack that utilizes a trade-secret coating that helps prevent PA from leeching out of the membrane. Metha Energy Solutions, in cooperation with Serenergy, revealed a hybrid electric/fuel cell vehicle in December of 2009 [101]. In this system, a methanol reformer is used to provide hydrogen to the PBI-based fuel cell. It was advertised that this vehicle could travel up to 310 miles on one tank of gas and takes only 2 min to refuel. EnerFuel, a subsidiary of Ener1, has also recently produced a hybrid electric/fuel cell vehicle. The EnerFuel EV uses a reformed methanol PBI fuel cell that works in conjunction with a lithium ion battery. The lithium battery is used to start the vehicle and to power the vehicle while driving, while the fuel cell system produces 3–5 kW to continuously recharge the battery. These fuel cell systems would not generate enough power to drive the vehicle, but would act as a range extender for the battery system. The target market of the EnerFuel EVs is not for those who drive 200+ miles daily, but instead for those with short daily commutes.

In July of 2009, the German Aerospace Center demonstrated that fuel cells have the potential of powering air-transportation vehicles [102, 103]. Designed in cooperation with Lange Aviation, BASF Fuel Cell, DLR Institute for Technical Thermodynamics, and



Serenergy, the Antares DLR-H<sub>2</sub> became the world's first piloted aircraft with a propulsion system powered only by PBI-based fuel cells. Besides creating zero CO<sub>2</sub> emissions during flight, the aircraft also generates much less noise than other comparable motor gliders. Using a fuel cell stack capable of producing up to 25 kW, the Antares DLR-H<sub>2</sub> has a cruising range of 750 km (or 5 h) and can travel at speeds of up to 170 km h<sup>-1</sup>. Similar fuel cell systems could be coupled with current commercial and military aircrafts as auxiliary power units (APUs) to improve fuel efficiency.

### Portable

Microelectricalmechanical (MEM) systems utilizing methanol reformers and PBI fuel cells have been developed for portable use. These devices are generally used to generate power in the range of 5–50 W for laptops, communication systems, and global positioning systems. Compared to batteries that offer equivalent amounts of power, these micro-fuel cell systems are lighter, generate less waste, and are overall more cost effective. Similar to other reformed methanol/PBI fuel cell systems, these MEMS are a sustainable technology as they reduce the amount of greenhouse gases produced per unit of electricity generated.

UltraCell of Livermore, CA, is a well-known producer of PBI MEM fuel cell systems. Funded and field tested by the US Army, the UltraCell XX25 is capable of providing 25 W of continuous maximum power [104]. Depending on the size of the fuel cartridge, the device is capable of delivering 20 W of continuous power from 9 h to 25 days. The fuel cartridge weighs less than a pound and The XX25 MEM system has been shown to power radio gear, mobile computer systems, communication devices, and a variety of other electrical devices. The XX25 provides roughly 70% in weight savings when compared to a typical battery on a 72-h mission (1.24 kg without the cartridge), and is rugged enough to operate in extremely cold or hot environments. In addition, it meets OSHA standards for safe indoor and in-vehicle use. Similar to the XX25, the newly developed UltraCell XX55 is capable of generating 55 W of continuous power for up to 2 weeks using the largest fuel cartridge [105]. Only 0.36 kg heavier than the XX25, the XX55 has an optional battery module that can provide a peak power output of 85 W.

Similar to the XX25, it is a very rugged device that can be used essentially in any conditions.

Larger than the Ultracell devices, the relatively new Serenergy Serenus E-350 is a reformed methanol/fuel cell hybrid with an approximate mass of 11 kg. At nominal power levels, it is capable of producing approximately 350 W [106]. The device is fueled by a 60–40 methanol-deionized water mixture. It takes approximately 45 min to start up, at which point it consumes fuel at a rate of 0.45 L h<sup>-1</sup>.

### H<sub>2</sub> Pump

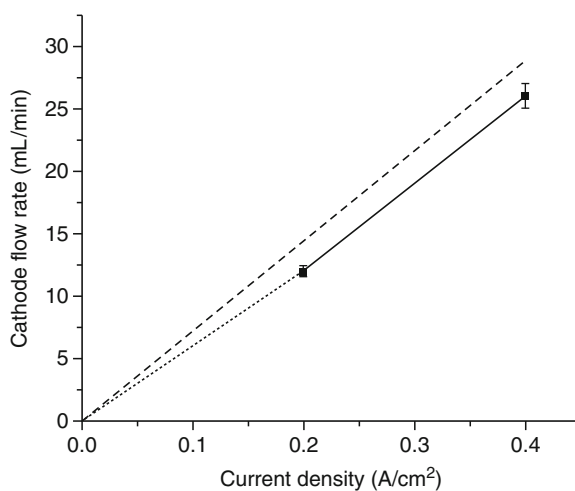
Efficient purification of hydrogen is becoming a common interest in both industrial and energy sectors. In particular, technology which can efficiently purify, pump, and pressurize hydrogen at low to moderate flow rates is needed, but is not readily available. Of course, there are existing methods for hydrogen purification which include various combinations of mechanical compression with cryogenic cleanup, palladium membranes, pressure swing absorption, and passive membrane separators to name a few. However, these technologies are challenged by certain limitations: (1) cryogenic cleanup produces high purity hydrogen, but requires costly refrigeration equipment and is suitable for very-large-scale specialty applications; (2) palladium membrane purification can be fairly simple in design and construction, but requires pressurization to drive the hydrogen separation process and suffers from poor utilization when purifying hydrogen from gases containing low fractions of hydrogen; and (3) pressure swing absorption (PSA) is widely used in high-volume industrial processes and relies on large, mechanical components that are subject to frequent maintenance and inherent inefficiency. Such devices are not easily scaled to smaller sizes or localized generation/purification needs. Furthermore, it is important to state that all of the above processes require expensive, high-maintenance compressors.

Electrochemical pumping is not a new concept and has in fact been utilized as a diagnostic technique within the electrochemical industry for years. General Electric developed this concept in the early 1970s [107]. The use of polymer electrolyte membranes for electrochemical hydrogen compression has been demonstrated in water electrolysis (H<sub>2</sub> generation) devices at

United Technologies Corporation, reaching 3,000 psi<sub>a</sub> [108], as well as studied in academic institutions [109]. The electrochemical hydrogen pump, first developed in the 1960s and 1970s, was derived from the original proton exchange membrane fuel cell efforts. The concept is simple, requires little power, and has been shown to pump hydrogen to high pressures. In the original work, the membrane transport medium was a perfluorosulfonic acid (PFSA) material, similar to the material used in many fuel cells today. The process is quite elegant in that like a fuel cell, molecular hydrogen enters the anode compartment, is oxidized to protons and electrons at the catalyst, and then the protons are driven through the membrane while the electrons are driven through the electrically conductive elements of the cell. The major difference in this cell as compared to a fuel cell is that the pump is operated in an electrolytic mode, not galvanic, meaning that power is required to “drive” the proton movement. Once the protons emerge from the membrane at the cathode, they recombine to form molecular hydrogen. Thus, hydrogen can be pumped and purified in a single step with a nonmechanical device. The pump concept builds upon the understanding of proton-transport membranes.

Clearly, the proton conducting membrane properties are critical. Desirable properties include: high proton conductivity, mechanical stability, low solubility and permeability of impurity gases, and sufficient operating temperature to support tolerance to impurities (CO and H<sub>2</sub>S) found in reformed gases. The application of the PBI membrane to electrochemical hydrogen pumping provides high proton conductivity (0.2–0.4 S cm<sup>-1</sup>), mechanical stability, enhanced gas separation, and up to 180°C operation. The high operating temperature eliminates water management difficulties typically experienced with the low operating temperatures of PSFA membranes while also providing tolerance to poisonous gas species such as CO. As such, the PBI membrane and electrode assembly represents a significant new opportunity and paradigm shift in electrochemical hydrogen pumps as well as in advancing the science of hydrogen separation, purification, and pressurization. This concept has been evaluated and demonstrated in recent work using PBI membranes [110]. The hydrogen pump was shown to operate with fairly low power requirements, and generally needed

less than 100 mV when operating at 0.2–0.4 A cm<sup>-2</sup>. This was accomplished without the critical water management commonly encountered in low-temperature, water-based membranes. The cathodic flow of hydrogen from the device was nearly identical to the theoretical Faradic flows. This suggests that the hydrogen pump could have applications as a hydrogen metering device since the hydrogen flow could be easily and accurately controlled by the current of the power source. The initial work reported devices that could operate for several thousand hours with little change in the operating parameters. This would be expected from the related work on PBI membranes for fuel cells which show outstanding long-term durability. In fuel cell applications, the ability to operate at high temperatures provides benefits for gas cleanup and durability on reformed fuels. In hydrogen pump applications, this tolerance to fuel impurities enables the hydrogen pump to purify hydrogen from hydrogen gas feeds containing such impurities. Figure 19 shows the operation of a PBI-based hydrogen pump operating on pure



**Polybenzimidazole Fuel Cell Technology. Figure 19** The cathodic flow rates of a hydrogen pump operated at 160°C and 0% relative humidity and fueled by pure hydrogen (*unfilled squares*), a reformat gas comprised of 35.8% H<sub>2</sub>, 11.9% CO<sub>2</sub>, 1,906 ppm CO, and 52.11% N<sub>2</sub> (*filled circles*), and a reformat gas comprised of 69.17% H<sub>2</sub>, 29.8% CO<sub>2</sub>, and 1.03% CO (*filled triangles*). The values are nearly identical, and thus, the symbols appear superimposed. The *dotted line* represents the theoretical flow rate at 100% efficiency [110]

hydrogen, as well as two different synthetic reformates. The flow rates are nearly unaffected by the composition of the gas feed at the various operating conditions (the data points are superimposed for the different gases). The data demonstrates that the pump was capable of operating at high CO levels (1% in this work) and extracting hydrogen from dilute feed streams (<40% hydrogen). Additionally, the hydrogen pump was capable of producing hydrogen with purities greater than 99%, with the final purity dependent on operating conditions. This device could play a prominent role for both the current industrial hydrogen users, as well as in a future economy that is more heavily reliant on hydrogen as an energy carrier. Commercial development of this device is underway.

### Conclusions and Future Directions

After approximately 10 years of development, PBI chemistries and the concomitant manufacturing processes have evolved to produce commercially available MEAs. PBI MEAs can operate reliably without complex water humidification hardware and are able to run at elevated temperatures of 120–180°C due to the physical and chemical robustness of PBI membranes. These higher temperatures improve the electrode kinetics and conductivity of the MEAs, simplify the water and thermal management of the systems, and significantly increase their tolerance to fuel impurities. Membranes cast by a newly developed PPA Process possessed excellent mechanical properties, higher PA/PBI ratios, and enhanced proton conductivities as compared to previous methods of membrane preparation.

The robustness of *p*-PBI membranes cast by the PPA Process has been tested and characterized by a variety of methods. Under a constant load, *p*-PBI has been shown to perform for well over 2 years with very little reduction in performance. Using synthetic reformates, *p*-PBI MEAs have demonstrated excellent resistances to impurities such as CO, CO<sub>2</sub>, and SO<sub>2</sub>. *p*-PBI membranes have also been shown to retain PA extremely well, and evidence strongly suggests that this small rate of PA loss would not significantly influence the life span of a MEA. Load, thermal, and shutdown–startup cycling tests of *p*-PBI fuel cells have also indicated comparable or improved results over other commercially available fuel cell systems. *p*-PBI is the

most common polymer in PBI-based fuel cell systems, although AB-PBI and other derivatives have been investigated. Recently developed 2OH-PBI, which to date has the highest recorded proton conductivity of all other PBIs, offers much potential for future fuel cell use.

Many fuel cell manufacturers are now considering the benefits of high-temperature PBI fuel cells. BASF Fuel Cell, the largest producer of PBI MEAs, has been in operation since March of 2007. BASF offers a wide variety of MEAs for stationary systems (combined heat and power, backup generators, etc.) and portable systems (transportation, microelectromechanical systems, etc.). Other companies, such as Plug Power, Serenergy, ClearEdge Power, and UltraCell, incorporate commercially available MEAs into their commercial fuel cell systems. Recently, H<sub>2</sub> Pump LLC has developed electrochemical pumping devices that use PBI membranes for the purification of hydrogen gas. Using various reformate gases, the devices have been shown capable of operating at high gas contamination levels and low hydrogen concentrations. Depending on operating conditions, the purity of the extracted hydrogen gas can be greater than 99%. In transportation applications, PBI-based fuel cells show great promise as APUs or range extenders for battery-powered electric vehicles.

### Acknowledgments

The authors would like to acknowledge BASF Fuel Cell GmbH and Dr. Gordon Calundann for technical and financial support of the work at Rensselaer Polytechnic Institute and at the University of South Carolina. The DOE-BES program [DEFG0205ER46258] and DARPA are also acknowledged for partial support in the preparation of the manuscript.

### Bibliography

1. Energy Information Administration (2008) World primary energy production by source (2008). Annual Energy Review. [http://www.eia.doe.gov/emeu/aer/pdf/pages/sec11\\_2.pdf](http://www.eia.doe.gov/emeu/aer/pdf/pages/sec11_2.pdf). Accessed 1 June 2010
2. U.S. Energy Information Administration (2008) Primary energy consumption by source and sector. Annual Energy Review. [http://www.eia.doe.gov/aer/pecss\\_diagram.html](http://www.eia.doe.gov/aer/pecss_diagram.html). Accessed 1 June 2010
3. U.S. Energy Information Administration (2008) Carbon dioxide emissions from the consumption and flaring of fossil fuels. Annual Energy Review. <http://www.eia.doe.gov/iea/overview.html>. Accessed 1 June 2010

4. Wainright J, Wang J, Weng D, Savinell R, Litt M (1995) Acid-doped polybenzimidazoles: a new polymer electrolyte. *J Electrochem Soc* 142(7):L121–L123
5. Dippel T, Kreuer KD, Lassegues JC, Rodriguez D (1993) Proton conductivity in fused phosphoric acid; a  $^1\text{H}/^{31}\text{P}$  PFG-NMR and QNS study. *Solid State Ionics* 61:41–46
6. Bozkurt A, Ise M, Kreuer KD, Meyer WH, Wegner G (1999) Proton-conducting polymer electrolytes based on phosphoric acid. *Solid State Ionics* 125:225–233
7. Weber J, Kreuer K-D, Maier J, Thomas A (2008) Proton conductivity enhancement by nanostructural control of poly(benzimidazole)-phosphoric acid adducts. *Adv Mater* 20:2595–2598
8. Vilciauskas L, Paddison SJ, Kreuer K-D (2009) Ab initio modeling of proton transfer in phosphoric acid clusters. *J Phys Chem A* 113:9193–9201
9. Mader J, Lixiang X, Schmidt T, Benicewicz BC (2008) Polybenzimidazole/acid complexes as high-temperature membranes. In: Scherer GG (ed) *Fuel cells II. Advances in Polymer Science*, vol 216. Springer, Heidelberg, pp 63–124
10. Wainright JS, Savinell RF, Litt MH (2003) High temperature membranes. In: Vielstich A, Lamm H, Gasteiger A (eds) *Handbook of fuel cells*, vol 3. Wiley, New York, pp 436–446
11. Colombari P (1992) Proton conductors: solids, membranes and gels – materials and devices. Cambridge University Press, Cambridge
12. Xiao L, Zhang H, Scanlon E, Ramanathan LS, Choe EW, Rogers D, Apple T, Benicewicz BC (2005) High-temperature polybenzimidazole fuel cell membranes via a sol-gel process. *Chem Mater* 17(21):5328–5333
13. Jayakody JRP, Chung SH, Durantino L, Zhang H, Xiao L, Benicewicz BC, Greenbaum SG (2007) NMR studies of mass transport in high-acid-content fuel cell membranes based on phosphoric acid and polybenzimidazole. *J Electrochem Soc* 154(2):B242–B246
14. Seel DC, Benicewicz BC, Xiao L, Schmidt TJ (2009) High-temperature polybenzimidazole-based membranes. In: Vielstich W, Gasteiger HA, Yokikawa H (eds) *Handbook of fuel cells*, vol 5. Wiley, Chichester, pp 300–312
15. Litt M, Ameri R, Wang Y, Savinell RF, Wainright JS (1999) Polybenzimidazoles/phosphoric acid solid polymer electrolytes: mechanical and electrical properties. *Mater Res Soc Symp Proc* 1999:313–323
16. Zhang H (2004) Novel phosphoric acid doped polybenzimidazole membranes for fuel cells. PhD Thesis, Rensselaer Polytechnic Institute, Troy, NY, December 2004
17. Li Q, Hjuler HA, Bjerrum NJ (2001) Phosphoric acid doped polybenzimidazole membranes: physicochemical characterization and fuel cell applications. *J Appl Electrochem* 31(7):773–779
18. Zhai Y, Zhang H, Liu G, Hu J, Yi B (2006) Performance degradation studies on PBI/ $\text{H}_3\text{PO}_4$  high temperature PEMFC and one-dimensional numerical analysis. *Electrochim Acta* 52(2):394–401
19. Kongstein OE, Berning T, Borresen B, Seland F, Tunold R (2007) Polymer electrolyte fuel cells based on phosphoric acid doped polybenzimidazole (PBI) membranes. *Energy* 32(4):418–422
20. Kim H, Cho SY, An SJ, Eun YC, Kim J, Yoon H, Kweon H, YEW KH (2004) Synthesis of poly(2, 5-benzimidazole) for use as a fuel-cell membrane. *Macromol Rapid Commun* 25(8):894–897
21. Asensio JA, Gomez-Romero P (2005) Recent developments on proton conducting poly(2, 5-benzimidazole) (AB-PBI) membranes for high temperature polymer electrolyte membrane fuel cells. *Fuel Cells* 5(3):336–343
22. Li QF, Jensen OJ (2008) Membranes for high temperature PEMFC based on acid-doped polybenzimidazoles. *Membr Technol* 2:61–96
23. Asensio JA, Borros S, Gomez-Romero P (2004) Polymer electrolyte fuel cells based on phosphoric acid-impregnated poly(2, 5-benzimidazole). *J Electrochem Soc* 151(2):A304–A310
24. Chen R (2003) Unpublished work
25. Yu S (2006) Novel polybenzimidazole derivatives for high temperature PEM fuel cells. PhD Thesis, Rensselaer Polytechnic Institute, Troy, NY
26. Delano CB, Doyle RR, Miligan RJ (1975) United States air force materials laboratory. Technical report, pp 1974–2022
27. Yu S, Zhang H, Xiao L, Choe EW, Benicewicz BC (2009) Synthesis of poly(2, 2'-(1, 4-phenylene) 5, 5'-bibenzimidazole) (para-PBI) and phosphoric acid doped membrane for fuel cells. *Fuel Cells* 9(4):318–324
28. Xiao L (2003) Novel polybenzimidazole derivatives for high temperature polymer electrolyte membrane fuel cell application. PhD Thesis, Rensselaer Polytechnic Institute, Troy, NY
29. Xiao L, Zhang H, Jana T, Scanlon E, Chen R, Choe EW, Ramanathan LS, Yu S, Benicewicz BC (2005) Synthesis and characterization of pyridine-based polybenzimidazoles for high temperature polymer electrolyte membrane fuel cell applications. *Fuel Cells* 5(2):287–295
30. Xiao L, Zhang H, Jana T, Scanlon E, Chen R, Choe EW, Ramanathan LS, Yu S, Benicewicz BC (2004) Synthesis and characterization of pyridine-based polybenzimidazoles for high temperature polymer electrolyte membrane fuel cell applications. *Fuel Cells* 5(2):287–295
31. Kallitsis JK, Gourdoupi N (2003) Proton conducting membranes based on polymer blends for use in high temperature PEM fuel cells. *J New Mat Electrochem Syst* 6(4):217–222
32. Faure S, Mercier R, Aldebert P, Pineri M, Sillion B (1996) Gas separation polyimide membranes used to prepare ion-exchange membranes for use in manufacture of fuel cells. French Pat 9605707
33. Watari T, Fang J, Tanaka K, Kita H, Okamoto KI, Hirano T (2004) Synthesis, water stability and proton conductivity of novel sulfonated polyimides from 4, 4'-bis(4-aminophenoxy) biphenyl-3, 3'-disulfonic acid. *J Membr Sci* 230(1–2):111–120
34. Lufano F, Gatto I, Staiti P, Antonucci V, Passalacqua E (2001) Sulfonated polysulfone ionomer membranes for fuel cells. *Solid State Ionics* 145(1–4):47–51
35. Einsla BR, Harrison WL, Tchatchoua C, McGrath JE (2004) Disulfonated polybenzoxazoles for proton exchange membrane fuel cell applications. *Polym Prepr* 44(2):645–646

36. Gil M, Ji X, Li X, Na H, Hampsey JE, Lu Y (2004) Direct synthesis of sulfonated aromatic poly(ether ether ketone) proton exchange membranes for fuel cell applications. *J Membr Sci* 234(1–2):75–81
37. Xing P, Robertson GP, Guiver MD, Mikhailenko SD, Kaliaguine S (2004) Sulfonated poly(aryl ether ketone)s containing the hexafluoroisopropylidene diphenyl moiety prepared by direct copolymerization, as proton exchange membranes for fuel cell application. *Macromolecules* 37(21):7960–7967
38. Gao Y, Robertson GP, Guiver MD, Mikhailenko SD, Kaliaguine S (2004) Synthesis of copoly(aryl ether ether nitrile)s containing sulfonic acid groups for PEM applications. *Macromolecules* 37(18):6748–6754
39. Xiao GY, Sun GM, Yan DY, Zhu PF, Tao P (2002) Synthesis of sulfonated poly(phthalazinone ether sulfones) by direct polymerization. *Polym Prepr* 43(19):5335–5339
40. Wang F, Hickner M, Kim YS, Zawodzinski TA, McGrath JE (2002) Direct polymerization of sulfonated poly(arylene ether sulfone) random (statistical) copolymers: candidates for new proton exchange membranes. *J Membr Sci* 197(1–2):231–242
41. Hickner MA, Ghassemi H, Kim YS, Einsla BR, McGrath JE (2004) Alternative polymer systems for proton exchange membranes (PEMs). *Chem Rev* 104(10):4587–4611
42. Kim S, Cameron DA, Lee Y, Reynolds JR, Savage CR (1996) Aromatic and rigid rod polyelectrolytes based on sulfonated poly(benzobisthiazoles). *J Polym Sci A* 34(3):481–492
43. Ariza MJ, Jones DJ, Roziere J (2002) Role of post-sulfonation thermal treatment in conducting and thermal properties of sulfuric acid sulfonated poly(benzimidazole) membranes. *Desalination* 147(1–3):183–189
44. Staiti P, Lufano F, Arico AS, Passalacqua E, Antonucci V (2001) Sulfonated polybenzimidazole membranes – preparation and physico-chemical characterization. *J Membr Sci* 188(1):71–78
45. Bae JM, Honma I, Murata M, Yamamoto T, Rikukawa M, Ogata N (2002) Properties of selected sulfonated polymers as proton-conducting electrolytes for polymer electrolyte fuel cells. *Solid State Ionics* 147(1–2):189–194
46. Asensio JA, Borros S, Gomez-Romero P (2002) Proton-conducting polymers based on benzimidazoles and sulfonated benzimidazoles. *J Polym Sci A* 40(21):3703–3710
47. Sakaguchi Y, Kitamura K, Nakao J, Hamamoto S, Tachimori H, Takase S (2001) Preparation and properties of sulfonated or phosphonated polybenzimidazoles and polybenzoxazoles. *Polym Mater Sci Eng* 84:899–900
48. Mader J (2010) Novel sulfonated polybenzimidazole derivatives for high temperature fuel cell applications. PhD Thesis, Rensselaer Polytechnic Institute, Troy, NY
49. He R, Li Q, Xiao GY, Bjerrum NJ (2003) Proton conductivity of phosphoric acid doped polybenzimidazole and its composites with inorganic proton conductors. *J Membr Sci* 226(1–2):169–184
50. Qian G, Benicewicz BC (2009) Synthesis and characterization of high molecular weight hexafluoroisopropylidene-containing polybenzimidazole for high-temperature polymer electrolyte membrane fuel cells. *J Polym Sci A* 47(16):4064–4073
51. Yu S, Benicewicz BC (2009) Synthesis and properties of functionalized polybenzimidazoles for high-temperature PEMFCs. *Macromolecules* 42(22):8640–8648
52. Scanlon E (2005) Polybenzimidazole based segmented block copolymers for high temperature fuel cell applications. PhD Thesis, Rensselaer Polytechnic Institute, Troy, NY
53. Schmidt TJ (2006) Durability and degradation in high-temperature polymer electrolyte fuel cells. *ECS Trans* 1(8):19–31
54. Schmidt TJ (2009) High-temperature polymer electrolyte fuel cells: durability insights. In: Buchi FN, Inaba M, Schmidt TJ (eds) *Polymer electrolyte fuel cell durability*. Springer, New York, pp 199–221
55. Ross PN Jr (1987) *Deactivation and poisoning of fuel cell catalysts*. Marcel Dekker, New York
56. Ferreira PJ, la O' GJ, Shao-Horn Y, Morgan D, Makharia R, Kocha S, Gasteiger HA (2005) Instability of Pt/C electrocatalysts in proton exchange membrane fuel cells: a mechanistic investigation. *J Electrochem Soc* 152(11):A2256–A2271
57. Tang L, Han B, Persson K, Friesen C, He T, Sieradzki K, Ceder G (2010) Electrochemical stability of nanometer-scale Pt particles in acidic environments. *J Am Chem Soc* 132(2):596–600
58. Liu G, Zhang H, Zhai Y, Zhang Y, Xu D, Shao Z-g (2007) Pt<sub>4</sub>ZrO<sub>2</sub>/C cathode catalyst for improved durability in high temperature PEMFC based on H<sub>3</sub>PO<sub>4</sub> doped PBI. *Electrochem Commun* 9(1):135–141
59. Landsman DA, Luczak FJ (2003) In: Vielstich W, Lamm A, Gasteiger H (eds) *Handbook of fuel cells – fundamentals, technology and applications*, vol 4. Wiley, Chichester, pp 811–831
60. Neyerlin KC, Singh A, Chu D (2008) Kinetic characterization of a Pt-Ni/C catalyst with a phosphoric acid doped PBI membrane in a proton exchange membrane fuel cell. *J Power Sources* 176(1):112–117
61. Schmidt TJ, Baurmeister J (2008) Properties of high-temperature PEFC Celtec-P 1000 MEAs in start/stop operation mode. *J Power Sources* 176(2):428–434
62. Luczak FJ, Landsman DA (1984) Ordered ternary fuel cell catalysts containing platinum, cobalt and chromium. US Patent 4,447,506
63. Luczak FJ, Landsman DA (1987) Ordered ternary fuel cell catalysts containing platinum and cobalt and method for making the catalyst. US Patent 4,677,092
64. Beard B, Ross PN Jr (1990) The structure and activity of platinum-cobalt alloys as oxygen reduction electrocatalysts. *J Electrochem Soc* 137(11):3368–3374
65. Glass JT, Cahen GL, Stoner GE (1987) The effect of metallurgical variables on the electrocatalytic properties of platinum-chromium alloys. *J Electrochem Soc* 134(1):58–65
66. Mukerjee S, Srinivasan S (1993) Enhanced electrocatalysis of oxygen reduction on platinum alloys in proton exchange membrane fuel cells. *J Electroanal Chem* 357(1–2):201–224
67. Mukerjee S, Srinivasan S, Soriaga MP (1995) Role of structural and electronic properties of Pt and Pt alloys on electrocatalysis of oxygen reduction. An in situ XANES and EXAFS investigation. *J Electrochem Soc* 142(5):1409–1422

68. Paulus UA, Scherer GG, Wokaun A, Schmidt TJ, Stamenkovic V, Radmilovic V, Markovic NM, Ross PN (2001) Oxygen reduction on carbon-supported Pt-Ni and Pt-Co alloy catalysts. *J Phys Chem B* 106(16):4181–4191
69. Stamenkovic V, Schmidt TJ, Ross PN, Markovic NM (2002) Surface composition effects in electrocatalysis: kinetics of oxygen reduction on well-defined Pt<sub>3</sub>Ni and Pt<sub>3</sub>Co alloy surfaces. *J Phys Chem B* 106(46):11970–11979
70. Stamenkovic V, Fowler B, Mun BS, Wang B, Ross PN, Lucas CA, Markovic NM (2007) Improved oxygen reduction activity on Pt<sub>3</sub>Ni(111) via increased surface site availability. *Science* 315(5811):493–497
71. Parrondo J, Mijangos F, Rambabu B (2010) Platinum/tin oxide/carbon cathode catalyst for high temperature PEM fuel cell. *J Power Sources* 195(13):3977–3983
72. Qian G (2008) Fluorine-containing polybenzimidazoles for high temperature polymer electrolyte membrane fuel cell applications. PhD Thesis, Rensselaer Polytechnic Institute, Troy, NY
73. Gasteiger H, Markovic NM (2009) Just a dream – or future reality? *Science* 324(5923):48–49
74. Debe MK, Schmoeckel AK, Vernstrom GD, Atanasoski R (2006) High voltage stability of nanostructured thin film catalysts for PEM fuel cells. *J Power Sources* 161(2):1002–1011
75. Strasser P (2009) Dealloyed Pt bimetallic electrocatalysts for oxygen reduction. In: Vielstich W, Yokokawa H, Gasteiger HA (eds) *Handbook of fuel cells: advances in electrocatalysis, materials, diagnostics, and durability*. Wiley, New York, pp 30–47
76. Lefevre M, Proietti E, Jaouen F, Dodelet J-P (2009) Iron-based catalysts with improved oxygen reduction activity in polymer electrolyte fuel cells. *Science* 324(5923):71–74
77. US Department of Energy (2010) MCFC and PAFC R&D workshop summary report. [http://www1.eere.energy.gov/hydrogenandfuelcells/pdfs/mcfc\\_pafc\\_workshop\\_summary.pdf](http://www1.eere.energy.gov/hydrogenandfuelcells/pdfs/mcfc_pafc_workshop_summary.pdf)
78. Yu S, Xiao L, Benicewicz BC (2008) Durability studies of PBI-based high temperature PEMFC. *Fuel Cells* 8(3–4):165–174
79. Okae I, Kato S, Seya A, Kamoshita T (1990) Study of the phosphoric acid management in PAFCs. In: *The Chemical Society of Japan 67th Spring Meeting*, vol 148, FCDIC Fuel Cell Symposium Proceedings, Japan
80. U.S. Department of Energy (2005) Manufacturing for the hydrogen economy: manufacturing research and development of PEM fuel cell systems for transportation application. [http://www1.eere.energy.gov/hydrogenandfuelcells/pdfs/mfg\\_wkshp\\_fuelcell.pdf](http://www1.eere.energy.gov/hydrogenandfuelcells/pdfs/mfg_wkshp_fuelcell.pdf)
81. California Fuel Cell Partnership (2009) Hydrogen fuel cell vehicle and station deployment plan: a strategy for meeting the challenge ahead. <http://www.fuelcellpartnership.org/>
82. Feitelberg AS, Stathopoulos J, Qi Z, Smith C, Elter JF (2005) Reliability of plug power GenSys fuel cell systems. *J Power Sources* 147(1–2):203–207
83. Schmidt TJ, Baurmeister J (2006) Durability and reliability in high temperature reformed hydrogen PEFCs. *ECS Trans* 3(1):861–869
84. Mocoteguy P, Ludwig B, Scholta J, Barrera R, Ginocchio S (2009) Long term testing in continuous mode of HT-PEMFC based H<sub>3</sub>PO<sub>4</sub>/PBI Celtec-P MEAs for u-CHP applications. *Fuel Cells* 9(4):325–348
85. Mocoteguy P, Ludwig B, Scholta J, Nedellec Y, Jones DJ, Roziere J (2010) Long-term testing in dynamic mode of HT-PEMFC H<sub>3</sub>PO<sub>4</sub>/PBI Celtec-P based membrane electrode assemblies for micro-CHP applications. *Fuel Cells* 10(2):299–311
86. Garsany Y, Gould BD, Baturina OA, Swider-Lyons KE (2009) Comparison of the sulfur poisoning of PBI and Nafion PEMFC cathodes. *Electrochem Solid State Lett* 12(9):B138–B140
87. Schmidt TJ, Baurmeister J (2008) Development status of PBI based high-temperature membrane electrode assemblies. *ECS Trans* 16(2):263–270
88. Reiser CA, Bregoli L, Patterson TW, Yi JS, Yang JD, Perry ML, Jarvi TD (2005) A reverse-current decay mechanism for fuel cells. *Electrochem Solid State Lett* 8(6):A273–A276
89. Puffer RH Jr, Rock SJ (2009) Recent advances in high temperature proton exchange membrane fuel cell manufacturing. *J Fuel Cell Sci Technol* 6(4):041013/1–041013/7
90. Puffer RH Jr, Hoppes GH (2004) Development of a flexible pilot high temperature MEA manufacturing line. *J Fuel Cell Sci Technol* 2004:573–579
91. Harris TAL, Walczyk D (2006) Development of a casting technique for membrane material used in high-temperature PEM fuel cells. *J Manuf Processes* 8(1):19–31
92. Birnbaum U, Haines M, Hake J-Fr, Linssen J (2008) Reduction of greenhouse gas emissions through fuel cell combined heat and power applications. 17th World Hydrogen Energy Conference, Forschungszentrum Juelich GmbH, IEA Greenhouse Gas R&D Programme, [www.fuelcells.org/info/residentialsavings.pdf](http://www.fuelcells.org/info/residentialsavings.pdf)
93. Schmitz R (2009) Japan working toward fuel-cell reality. *Marketplace*, 8 Dec 2009, p 1
94. BASF Fuel Cell GmbH (2010) Celtec MEAs: membrane electrode assemblies for high temperature PEM fuel cells. <http://www.basf-fuelcell.com/en/projects/celtec-mea.html>
95. Plug Power (2009) High-temperature fuel cell system for residential applications. <http://www.plugpower.com/products/residentialgensys/residentialgensys.aspx>
96. Serenergy (2010) Serenus 166/390 Air C v2.5. [http://www.serenergy.com/files/assets/documentation/166\\_390%20Air%20C%20v2.5%20data%20sheet\\_v1.1-0210.pdf](http://www.serenergy.com/files/assets/documentation/166_390%20Air%20C%20v2.5%20data%20sheet_v1.1-0210.pdf)
97. ClearEdge Power (2009) Delivering smart energy today. <http://www.clearedgepower.com/>
98. U.S. Energy Information Administration (2009) Emissions of greenhouse gases report. <http://www.eia.doe.gov/oiaf/1605/ggrpt/#ercde>
99. Wired.com (2010) A hydrogen highway for the east coast. <http://www.wired.com/autopia/2010/01/east-coast-hydrogen-highway/>
100. Telegraph.co.uk, Telegraph Media Group Limited (2008) Driving VW's fuel-cell prototypes. <http://www.telegraph.co.uk/motoring/green-motoring/3520714/Driving-VWs-fuel-cell-prototypes.html>

101. Renewableenergyfocus.com (2009) Innovative Danish technology uses methanol to make fuel cell vehicles competitive, December 2009. <http://www.renewableenergyfocus.com/view/5650/innovative-danish-technology-uses-methanol-to-make-fuel-cell-vehicles-competitive/>
102. German Aerospace Center (2009) DLR motor glider antares takes off in Hamburg – powered by a fuel cell. [http://www.dlr.de/en/desktopdefault.aspx/tabid-13/135\\_read-18278/](http://www.dlr.de/en/desktopdefault.aspx/tabid-13/135_read-18278/)
103. What's next after the dreamliner? Think fuel cells. Design News, 18 Sept 2009. [http://www.designnews.com/article/354516-What\\_s\\_Next\\_after\\_the\\_Dreamliner\\_Think\\_Fuel\\_Cells.php](http://www.designnews.com/article/354516-What_s_Next_after_the_Dreamliner_Think_Fuel_Cells.php)
104. Ultracell (2008) UltraCell XX25: mobile power for mobile applications. [http://www.ultracellpower.com/assets/XX25\\_Data\\_Sheet\\_01-22-2009.pdf](http://www.ultracellpower.com/assets/XX25_Data_Sheet_01-22-2009.pdf)
105. Ultracell (2008) UltraCell XX55: extreme mobile power for demanding applications. [http://www.ultracellpower.com/assets/XX55\\_Data\\_Sheet\\_01-27-2009.pdf](http://www.ultracellpower.com/assets/XX55_Data_Sheet_01-27-2009.pdf)
106. Serenergy (2010) Serenus methanol fuel cell module – 350W. [http://www.serenergy.com/files/assets/documentation/Serenus%20H3%20E-350\\_datasheet\\_v1.1-0210.pdf](http://www.serenergy.com/files/assets/documentation/Serenus%20H3%20E-350_datasheet_v1.1-0210.pdf)
107. Maget HJR (1970) Process for gas purification. US Patent 3489670, 13 Jan 1970
108. McElroy JF (1989) SPE regenerative hydrogen/oxygen fuel cells for extraterrestrial surface applications. In: Energy conversion engineering conference, Washington, DC, Aug 1989. IEEE, Washington, DC, pp 1631–1636
109. Rohland B, Eberle K, Strobel R, Scholta J, Garcke J (1998) Electrochemical hydrogen compressor. *Electrochim Acta* 43(24):3841
110. Perry KA, Eisman GA, Benicewicz BC (2008) Electrochemical hydrogen pumping using a high-temperature polybenzimidazole (PBI) membrane. *J Power Sources* 177(2): 478–484

## Polyculture in Aquaculture

ROBERT R. STICKNEY

Texas Sea Grant College Program, Texas A&M University, College Station, TX, USA

### Article Outline

Glossary  
 Definition of Polyculture  
 Introduction  
 Key Principles  
 Polyculture Successes  
 Future Directions  
 Bibliography

### Glossary

**Benthos** Organisms that live on or in the sediments in aquatic environments.

**Niche** A habitat that provides for the needs to support the life of an organism.

**Plankton** The community of plants and animals suspended in the water column that drift with the currents. They have limited or no swimming ability but may be able to move vertically.

**Phytoplankton** The plant component of the plankton community.

**Zooplankton** The animal component of the plankton community.

### Definition of Polyculture

Polyculture is the production of two or more cultured species in the same physical space at the same time, often with the objective of producing multiple products that have economic value. They may be a combination of animals, plants and animals, aquatic species only, or aquatic and terrestrial species.

### Introduction

Aquaculture has its roots in China, perhaps as long as 4,000 years ago [1]. Interestingly, polyculture was a part of that early history. The Chinese often stocked multiple species of carp together in ponds to take advantage of all the types of food available. The pond would be fertilized, often with terrestrial animal manure, to promote the growth of the food organisms. In addition, terrestrial vegetation might be added. Thus, each of the species stocked occupied its own ecological niche.

Another form of polyculture that has a long history is rice–fish culture. By modifying rice ponds to provide a refuge for fish when a rice paddy is dewatered (usually a trench down one side or in the middle of the paddy), the rice farmer can obtain two types of crops. The most common fishes employed in rice–fish farming are common carp (*Cyprinus carpio*) and tilapia (*Oreochromis* sp.).

Today, polyculture employs a much wider variety of species than those that have been used for millennia in China, and there have been nuances developed in the polyculture approach. In most cases, the species used in polyculture systems need to be compatible, that is, they

need to grow well without interfering with one another. One example is culturing freshwater shrimp (*Macrobrachium rosenbergii*) with tilapia [2]. Exceptions do occur. For example, a predatory species may be stocked to prey on unwanted offspring of the primary culture species. A good example is tilapia culture, where stocked fingerlings may become reproductively active well in advance of reaching desirable market size. A fish predator, such as the snakehead (*Channa* sp.), can be stocked at a size too small to consume the tilapia that are in the system for growout, but large enough to prey upon unwanted tilapia fry.

In marine cage culture, fouling by such organisms as bryozoans and barnacles is often a serious problem. Netting may become so fouled that circulation through the cage is reduced to the point that dissolved oxygen depletions can occur. Depending upon the location of the cage culture operation, such fouling can occur within several days to a few weeks, so frequent cleaning is required. If an animal that will consume the fouling organisms can be found and stocked in the cages, it may be possible to extend the time between manual cleanings. There have also been instances where fish have been stocked in marine cages to consume parasites. One example is the stocking of wrasse (family Labridae) to control sea lice in Atlantic salmon cages in Norway [3]. Another is stocking sea cucumbers (class Holothuroidea) in salmon net pens to feed on fish feces, fouling organisms, and unconsumed feed [4].

Polyculture is primarily used to enhance total production within an aquaculture facility while maintaining and, in many cases, enhancing water quality. Mussels grown in the vicinity of salmon net pens, for example, remove phytoplankton that take up nutrients associated with the degradation of fish feces and unconsumed feed. Adding a seaweed, such as kelp (order Laminariales), to the system can further reduce the levels of dissolved nitrogen and phosphorus in the water. Polyculture of fish, shellfish, and algae together is an example of integrated multitrophic aquaculture [5]. While Atlantic salmon (*Salmo salar*) is the most common fish cultured using the technique, it has also been used with rabbitfish (*Siganus* sp.) [6].

Another specialized case of polyculture is hydroponics (sometimes called aquaponics), wherein terrestrial plants are grown in a nutrient solution rather than

being planted in soil. Hydroponics is usually conducted in greenhouses. The water for the terrestrial plants, which are often a combination of vegetables and/or herbs, is fertilized to provide all the proper nutrients for rapid growth and, of course, provided with sufficient natural or artificial light for photosynthesis. In many cases, one or more aquatic species (fish or shellfish) are incorporated into the system.

Many nations culture fish in ponds that receive avian, livestock, or even human waste as fertilizer. The practice is particularly common in developing countries where inorganic fertilizer is expensive and often difficult to obtain. Common carp and tilapia are often grown in ponds that receive the waste from the terrestrial animals.

### Key Principles

#### Polyculture is a Technique that Produces Two or More Compatible Species Within the Same Culture System

Not every species in a polyculture system needs to be aquatic in nature. Combinations of terrestrial and aquatic species are common. One or more of the species in a polyculture system may be a terrestrial or an aquatic plant.

#### Polyculture Systems can Increase the Profitability of a Culture System

By culturing two or more marketable species together in the same culture system or in close proximity to one another, it is often possible to increase the amount of income generated by a producer. While it may not be economically possible to rear one of the species profitably alone, by polyculturing it with a high-value species, total revenue can be increased. An example is culturing seaweed that is a good source of agar or carrageenan and that could add value to a culture facility if grown in the vicinity of a higher value crop like salmon [7].

#### Polyculture Systems can Reduce the Environmental Impacts from Aquaculture Systems

Nutrients released from cultured fish and shellfish-culture facilities can be effectively removed from the water by culturing plants such as seaweeds in close proximity to the animals [7]. An alternative approach



would involve the nutrients from such species as fish or shrimp to support phytoplankton that could, in turn, provide a food source for filter feeding Mollusca, such as oysters, scallops, and mussels [5].

### Polyculture Successes

As mentioned, polyculture was developed 1000s of years ago [1] and must be considered successful without question since it is still being employed, not only in China, but also in many other countries. The objective of the Chinese system is to stock species that take advantage of the various available food sources in culture ponds. Silver carp (*Hypophthalmichthys molitrix*) consume phytoplankton, bighead carp (*Hypophthalmichthys nobilis*) graze on zooplankton, mud carp (*Cirrhinus molitorella*) feed primarily on detritus, black carp (*Mylopharyngodon piceus*) are mollusc eaters, while grass carp (*Ctenopharyngodon idella*) consume higher vegetation. Grass carp will eat some species of aquatic macrophytes and will also consume various types of terrestrial plants that may be added to the culture pond. Other species that are used in carp polyculture systems are common carp (*Cyprinus carpio*) and tilapia (*Oreochromis* spp.). There are a wide variety of fish, shellfish, and plants that are used in polyculture around the world. In many cases, they are local species, though introduced species are also commonly found – in many cases those were introduced decades ago, such as is the case with tilapia in Asia and the Americas.

Integrated multi-trophic aquaculture systems have a relatively short history, but appear to be highly effective and are a modern approach to polyculture that is being increasingly adopted. Thus, that approach can also be considered a success. The approach typically involves the co-culture of carnivorous species, such as finfish or shrimp, with a filter feeder and a primary producer, typically some kind of algae [8, 9].

Hydroponics can be considered a success, in that the approach can be used to produce a variety of vegetables and other types of plants. While capital intensive, hydroponic systems can be profitable. Various species of aquatic animals have been grown in such systems to provide nutrients that the plants can use and also to produce an additional marketable product. The

aquatic animals cannot be counted upon as the only or even the primary nutrient source for the plants, however. The nutrient levels in the aquatic-animal waste products are insufficient in volume and composition, so a nutrient broth needs to be provided.

### Future Directions

Biofloc aquaculture is an approach that has been around since the 1970s, but has as yet to be widely adopted. In biofloc systems, the development of high levels of nonpathogenic bacteria in the water is encouraged. Heavy aeration is required to maintain the dissolved oxygen level for the culture species, which is most commonly shrimp, though such systems have also been used in conjunction with tilapia and other fishes. As more experience and research has occurred over the years, the methods for developing and controlling water quality in such systems have advanced. In 2009, one research group was able to produce shrimp at a final density of 9.75 kg/m<sup>3</sup>, which is as much as ten times higher than what is typical [10]. It is likely that the approach will be more widely adopted in the future.

Open-ocean aquaculture is expanding rapidly around much of the globe. In the USA, marine aquaculture is largely confined to production of molluscs in coastal waters and salmon cage and net pen culture in protected waters. Once a regulatory framework for the United States Exclusive Economic Zone has been promulgated, there is a likelihood that offshore aquaculture facilities will be developed. It seems reasonable to assume that a polyculture approach similar to the one that is commonly seen with respect to salmon culture, particularly in Chile, will be employed in conjunction with offshore aquaculture as a means of helping prevent water-quality degradation and increasing profitability. Offshore polyculture may involve rearing molluscs suspended from platforms used as support facilities for sea-cage culture operations. Such support structures may provide housing for the culturists, feed, and equipment storage, and may also be used as hatcheries (oil and gas platforms being appropriate once they are out of production and which have been of interest to some prospective aquaculturists). Seaweeds could also be produced, so an offshore operation might employ the multitrophic approach to culture.

## Bibliography

1. Stickney RR (2009) *Aquaculture: an introductory text*. CABl, Wallingford, United Kingdom p 304
2. Brick RW, Stickney RR (1979) Polyculture of *Tilapia aurea* and *Macrobrachium rosenbergii* in Texas. *Proc World Maricult Soc* 10:222–228.
3. Sayer MDJ, Treasurer JW, Costello MJ (1996) *Wrasse biology and use in aquaculture*. Wiley, New York p 296.
4. Ahlgren MO (1998) Consumption and assimilation of salmon net pen fouling debris by the red sea cucumber *Parastichopus californicus*: implications for polyculture. *J World Aquacult Soc* 29:133–139
5. Chopin T, Bastarache S (2004) Mariculture in Canada: finfish, shellfish and seaweed. *World Aquacult* 34(3):37–41
6. Mmochi AJ, Mozes N, Kissil G, Dubi AM, Jiddawi NS, Mwangamilo J (2001) Design and preliminary results of an integrated mariculture pond system (IMPS) at Makoba, Zanzibar. *Tanzan Mar Sci Devel In Tanzan East Afr* 1:431–450
7. Troell M, Nilsson C, Buschmann A, Kautsky AH, Kautsky N (1997) Integrated marine cultivation of *Gracilaria chilensis* (Gracilariales, Bangiophyceae) and salmon cages for reduced environmental impact and increased economic output. *Aquacult* 156:45–61
8. Neori A, Troell M, Chopin T, Yarish C (2007) The need for a balanced approach to blue revolution aquaculture. *Environ* 49:37–43
9. Ridler N, Wowchuk M, Robinson B, Barreington K, Chopin T, Robinson S, Page F, Reid G, Szemerda M, Sewuster J, Boyne-Travis S (2007) Integrated multi-trophic aquaculture (IMTA): a potential strategic choice for farmers. *Aquacult Econ Manage* 11:99–110
10. Samochoa T, Correia E, Hanson T, Wilkenfeld J, Morris T (2010) Operation and economics of a biofloc-dominated zero exchange system for the production of Pacific white shrimp, *Litopenaeus vannamei*, in greenhouse-enclosed raceways. In: *Proceedings of the AES issues forum, Roanoke, 18–19 Aug 2010*

## Polymer Electrolyte (PE) Fuel Cell Systems

JOHN F. ELTER

Sustainable Systems, LLC, Latham, New York, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

The PEM Fuel Cell Industry

Fuel Cell Basics

Critical Parameters/Critical Specifications

Recent Advancements

Fuel Cell System Design

Technical Readiness

System Cost and Reliability

Fuel Cell Systems and Sustainability

Future Directions

Bibliography

### Glossary

**Electrocatalyst** A material that enhances the rate of an electrochemical reaction, such as the hydrogen oxidation reaction (HOR) or the oxygen reduction reaction (ORR), without itself being consumed in the reaction. In most PEM fuel cells, the electrocatalysts are nanosized materials made from the precious metals group (PGM), usually platinum, palladium, and ruthenium, or alloys of these materials with nickel, cobalt, and manganese.

**Fuel cell system (FCS)** The combination of a fuel cell stack and the subsystems that are required to support its operation for the intended application.

**Fuel cells (FC)** Electrochemical cells that convert chemical energy from a fuel into electrical energy through the controlled transfer of electrical charge driven by the difference in electrochemical potential between two electrodes separated by an electrolyte.

**Gas diffusion layer (GDL)** A component of the fuel cell used to evenly distribute reactants across the electrode surface.

**Membrane electrode assembly (MEA)** A manufactured unit consisting of the electrodes, membrane, and gas diffusion layers. Sometimes, the MEA can include the gasket used to seal the fuel cell. The MEA, when placed between two electrically conducting plates with channels to provide the fuel and oxygen, constitute a single cell.

**Polymer electrolyte membrane (PEM)** Cation- or anion-conducting polymer membrane that separates the two electrodes in a fuel cell. In most PEM fuel cells, the polymer is a cation (proton) conductor and in this case PEM can stand for proton exchange membrane.

**Stack** A collection of individual fuel cells, separated by electrically conducting “bipolar” plates, connected in series. Stacks can range in size from a few watts as in portable fuel cells to hundreds kilowatts as in stationary fuel cells for combined heat and power.

**Triple phase boundary (TPB)** The place on the catalyst surface within the fuel cell electrodes where protons, electrons, and the fuel meet in order to oxidize hydrogen or reduce oxygen reactants.

### Definition of the Subject

This chapter discusses the subject of fuel cells: how they work and how they are designed and integrated into a collection of subsystems for application in a variety of applications. Particular focus is placed upon systems utilizing polymer electrolyte membranes, and how the properties of the membrane dictate the system design considerations.

Fuel cells themselves were invented over 170 years ago, with the first fuel cell demonstrated by Sir William Robert Grove in 1839. The first commercial use of a PEM fuel cell in a system occurred when General Electric developed the technology with NASA and McDonnell Aircraft for use in the Gemini space program. Since then, fuel cells have been integrated into systems intended for a wide variety of applications, from fuel cells of just a few watts for portable application, to hundreds of kilowatts for applications in transportation.

### Introduction

The vision of fuel cells running on renewable hydrogen as an energy carrier derived from water, and producing electricity with only water as a by-product, has long captured the imagination of those concerned with the environment. Since the early days of the use of fuel cells in the Gemini space program, various explorations of the potential use of fuel cells have been conducted by individuals, small firms, and large corporations. In each case, the potential of fuel cells has been validated. The long-term benefits of high efficiency and fuel flexibility are clear, especially in today's context. Energy security and the need to reduce dependence on foreign oil, and the need for sustainable economic development and the reduction of greenhouse gas emissions have been brought into sharp focus by war, weather, and wildlife.

Great progress has been and continues to be made in the development of fuel cells since their first commercial application roughly 50 years ago. This is particularly true of the fuel cells developed for automotive applications, which represents a very demanding application in terms of durability and cost. Progress has been made in other applications as well, particularly for those being developed for motive power and stationary combined heat and power. This chapter will attempt to lay out the status of the technology today, and where it needs to go in the future. First, a brief overview will be given of the fuel cell industry, followed by a review of some of the basic principles underlying the theory of operation. The notion of critical design parameters and specifications will be discussed as a prelude to a review of some recent advancement and to the system design process and some important considerations for fuel cell systems. Finally, there will be a discussion on future directions.

### The PEM Fuel Cell Industry

In reality, it is fair to say that the fuel cell industry is nascent. The long-term benefits of high efficiency and potential fuel flexibility are clear, but cost and durability issues are causing a slow adoption timeline. In addition, the nagging issue of the infrastructure needed to provide hydrogen as a fuel has not gone away. Hydrogen availability is still a relevant issue for many of the applications that would have fuel cells as the privileged technology. Other issues include the operational flexibility and lifespan of the incumbent technology. In response to these important issues, fuel cell manufacturers, in need of revenues, are attempting to approach large niche markets for which the technological requirements are suited to current fuel cell capabilities. These markets, such as backup power, indoor forklift trucks, bus fleets, and consumer electronics, each has their own set of adoption issues. Even in these cases, government subsidies are needed to help the industry cope with the slow market penetration associated with product immaturity and the risk-adverse nature of the industries being targeted. Limited revenues from these markets lead to high stock volatility, making valuations difficult, limiting private investment and further exasperating the situation.

The fuel cell industry is itself highly fragmented. Investment bankers such as JP Morgan have limited their coverage to those companies with larger market caps, including Ballard and Plug Power, who focus on backup, forklift, and residential cogeneration systems, and for example Medis Technologies which focuses on the consumer electronics industry with a proprietary direct liquid fuel cell battery charging device. Other players in the fuel cell world include companies with a variety of business models that are aggressively pursuing similar markets. These include pure stack manufacturers, such as Nedstack, fuel cell systems providers such as ClearEdge and IdaTech, companies that supply both stacks and systems to end user and OEM customers, such as Nuvera, and companies that provide stack technology to OEMs such as Intelligent Energy. Feeding fuel cell components to these companies are much larger corporations such as 3M, DuPont, Gore, Asahi Glass, and BASF, as well as a host of large automotive suppliers providing the “balance of plant” components such as blowers, compressors, heat exchangers, and electronics.

One reason for the fragmentation of the fuel cell industry is the wide range of applications to which fuel cells can be applied. As mentioned, these include transportation applications (with the key players being General Motors, Toyota, Honda, UTC, and others), material handling (Ballard, Plug Power, Nuvera, Hydrogenics), telecommunications backup (Plug Power, Ballard, ReliOn, IdaTech), consumer electronics (Samsung, Toshiba, and others) and residential cogeneration (Plug Power, ClearEdge, Toshiba, others). Each of these markets has a well-defined set of operational characteristics and requirements, as defined by the customer. Another and related reason for the fragmentation of the industry is the lack of a dominant design paradigm. Each company has a preferred method of approaching the design of a fuel cell and the system into which it is embedded. Since no single approach has emerged as dominant, there is a lack of standards, de facto or otherwise, around which the industry can capture the benefits of economies of scale and of accumulated knowledge.

On the other hand, it might be argued that there is a de facto dominant design, and it is that standard which, up to now unwittingly adhered to, is limiting the true out-of-the-box thinking needed for step

changes in performance at lower cost and improved reliability. In most cases, as will be seen, the commercialization issues around each application are driven to a large extent by this underlying “standard” technology, and its modes of failure in meeting customer requirements. In this chapter, specific attention will be given to these failure modes, and how they drive system complexity. In addition, some attention will be paid to the system design, and how it can be simplified to enable lower costs.

Finally, as will be discussed here, there has been steady progress in lowering costs and improving reliability, due in part to sustained government support of fuel cells over the past decade. Still, there is “a lot of room at the bottom” for scientific and engineering innovation. Indeed, the application of nanotechnology as a means of enabling the fuel cell industry to become a sustainable ecosystem should not be underestimated. There is ample reason for considerable optimism around PEM fuel cell systems for the sustainable energy applications they support.

## Fuel Cell Basics

Fuel cells are electrochemical devices that transform chemical energy into electrical energy, much the same way a battery converts chemical potential energy into electrical kinetic energy, except that the chemicals consumed in a fuel cell in the reaction are supplied from external sources. The fuel cell enables the generation of electricity while producing pure water by controlling the combining of hydrogen and oxygen:



Ideally, this reaction would take place at a potential of 1.229 V at standard conditions (25°C and 1 atm) if were it not for losses due to irreversible internal processes. In fact, the reversible potential  $E_r$  of this reaction can be expressed in terms of the change in the Gibbs free energy,  $\Delta G$ :

$$\Delta G = -nFE_r \quad (2)$$

Here  $n$  is the number of electrons involved and  $F$  is Faraday’s constant, which has the value of 96,485 coulombs of charge per mole of electrons. The derivation of Eq. 2 is a straightforward application of thermodynamics, recognizing that in this case the work done is

the product of the amount of charge transferred and the potential. For the case of standard conditions, it is known from thermodynamics that  $\Delta G = 237$  kJ/mole when the water is produced as a liquid (corresponding to the higher heating value, HHV) and  $n = 2$  for the  $\text{H}_2/\text{O}_2$  fuel cell reaction. Of course, not all of the energy in the reaction is turned into electricity, and because of this some portion of hydrogen's HHV is converted into heat. The enthalpy change in the reaction is given by the familiar thermodynamic relationship

$$\Delta H = \Delta G + T\Delta S \quad (3)$$

From thermodynamic tables, it is found that at standard conditions  $\Delta H = 286$  kJ/mole. It follows that roughly 49 kJ/mole are converted into heat and the theoretical efficiency  $\eta_{\text{th}}$  of a fuel cell operating at standard conditions is

$$\eta_{\text{th}} = \Delta G/\Delta H = 237/286 = 83\% \quad (4)$$

There are various additional losses that are due to irreversible processes in the cell itself. These losses, which are intrinsic to the cell and are materials dependent, reduce the fuel cell efficiencies to around 40–60%, depending on the type of fuel cell and its application. There are numerous references that discuss the theory behind fuel cells in detail [1, 2], as well as the other chapters in this encyclopedia. Here only a very cursory overview is given, as the focus of this chapter is on the fuel cell *system*.

In a typical fuel cell, there are two electrodes, an anode from which electrons are removed during the oxidation of hydrogen, a cathode to which electrons flow during the reduction of oxygen, an ion-conducting electrolyte which electrochemically connects the anode and cathode, and an external circuit through which the electrons flow from the anode to the cathode while doing electrical work. In early fuel cells, the electrolyte was generally a liquid such as a solution of KOH, which provided for the flow of  $\text{OH}^-$  anions. In this case, the fuel cell was termed an “alkaline” fuel cell because of the nature of the electrolyte. In the same way, the electrolyte can be phosphoric acid, in this case providing for the flow of  $\text{H}^+$  cations, or protons. Such a fuel cell is called a “phosphoric acid” fuel cell, again in recognition of the electrolyte. In this chapter, polymer electrolyte fuel cell systems will be discussed. In this

case, the electrolyte is a polymer in the form of a *membrane*. Hence the term or “polymer electrolyte membrane” or PEM fuel cell. PEM fuel cells, or PEMFCs, are electrochemical devices whose characteristics are determined largely by the properties of the polymer electrolyte. One of the main purposes of this chapter is to illustrate the extent to which the PEMFC's performance is in fact *dominated* by the properties of the polymer electrolyte. It is worth noting that in some cases the acronym PEM is interpreted to mean Proton Exchange Membrane, since in most cases in which the acronym PEM is used, the ion that is transported across the membrane and “exchanged” at the electrodes is the proton. However, here the acronym will be interpreted to mean Polymer Electrolyte Membrane to emphasize the point that it is the polymer electrolyte that to a large extent dictates the characteristics of the system.

In a fuel cell, the two reactants, in this case hydrogen and oxygen, are each directed macroscopically by flow channels molded into electrically and thermally conductive plates. The hydrogen on the anode side and the oxygen at the cathode side are each brought into contact with an electrode by the gas diffusion layer or “GDL” which ensures uniform flow across the cell. As discussed, a polymer membrane is provided that serves as the electrolyte, providing good ionic conductivity while maintaining the separation of these reactant gases, and preventing the flow of electrons. The membrane needs to make intimate contact with the electrodes to ensure efficient charge transfer across the embedded electrode/electrolyte interface. The electrodes are generally porous to allow the diffusion of the reactant gases while providing a support for the catalysts needed to facilitate the electrochemical reactions. The electrodes also need to provide conductive pathways for the ions and electrons. Clearly, the polymer electrolyte membrane, electrodes, and gas diffusion layer have a complex set of interacting functional requirements. These will be discussed in detail, especially from the point of view of their impact on system design.

In most cases, the polymer electrolyte is basically acidic, in which case, the catalysts are usually based upon platinum and its alloys. At the anode electrode, hydrogen gas molecules adsorb onto the supported catalyst material which is used to help strip the

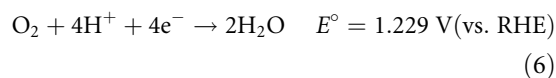
hydrogen of its electron, oxidizing it to become a hydrogen cation, or proton. This proton is transported through the polymer electrolyte membrane to the cathode electrode. Here, oxygen molecules are fed into the cathode electrode, usually by supplying air. Like the active layer at the anode, appropriate nanosized catalyst particles are supported on electrically conducting porous materials. In the current dominant design paradigm, this support is usually a high surface area carbon. Once on the catalyst surface, the oxygen's covalent bonds are broken, forming oxygen anions. The oxygen anions then combine with protons arriving through the membrane and electrons that have been driven through an external circuit by the difference in the anode and cathode potentials. The *critical point* on the catalyst surface at which protons, electrons, and reactant gases must come together is generally called the *triple phase boundary* (TPB). All of the necessary components, the gas diffusion media, the porous electrodes, and the polymer electrolyte membrane, and if desired, even the conformable gasket required to provide sealing around the perimeter of the plate, are manufactured as a “membrane electrode assembly” or “MEA” by various suppliers like 3M, Gore, OMG, and others. Since a typical PEMFC operates at a voltage of less than a volt, multiple cells are electrically connected in series to form a “stack” of cells that can be sized for a particular application. In this case, each thermally and electrically conducting plate is shared between the anode of one cell and the cathode of another, and is thus termed “bipolar.” The integrated stack design must accommodate a means for managing the products of the fuel cell reaction: electricity, water, and heat. In smaller systems, the stack can be air cooled, but in larger systems, the plates must incorporate channels that allow for the flow of a liquid coolant.

The electrochemical equation governing the hydrogen oxidation reaction (HOR) at the anode of a single cell is

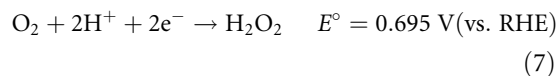


This is a thermodynamically reversible reaction, and with platinum as the catalyst, is a standard reference known as the “reversible hydrogen electrode” (RHE), for which the potential is universally chosen as zero volts. The oxygen reduction reaction (ORR) at

the cathode, on the other hand, is thermodynamically irreversible, and is generally expressed in terms of the dominant four electron reaction:



In fact, the ORR is an extremely complicated reaction that is dependent upon the electronic structure and number of the active sites on the catalyst surface, which in turn are influenced by many factors including the size and shape of the catalyst particle, the presence of any contaminant species competing for reaction sites, the presence of oxides which may be formed during potential cycling, the effect of the catalyst support, etc. In addition to the 4-electron reaction, there is the possibility for the 2-electron peroxide pathway [3],



In low-temperature PEMFCs, those that operate in the range of 65–80°C, the peroxide pathway is problematic because if the peroxide is not quickly decomposed into water, it can form radicals that chemically attack the membrane, causing premature failure.

The ORR is a major source of efficiency loss in a PEMFC, but it is not the only source of the so-called overpotential. In general, there are also resistive and mass transport losses. The resistive losses are associated with the finite conductivities of the electrolyte and the electrodes and the contact resistance losses at the plate and GDL interfaces, whereas mass transport losses are associated with the lack of adequate fuel or air reaching the reaction sites within the electrode. For low-temperature PEMFC systems that operate below the boiling point of water, mass transport losses are primarily due to the buildup of liquid water in the electrode or gas diffusion layer (referred to as “flooding”). Mass transport losses are also associated with the tortuous pathways that exist in the porous electrodes and the GDL. Reducing all of these sources of potential losses is important for achieving the highest possible efficiency.

It is perhaps convenient at this point to briefly mention the types of PEM fuel cell that are under development. Low-temperature PEM fuel cells (LTPEMFCs) operate in the range of 65–80°C. These

systems are limited in temperature by the requirement for the membrane to be maintained in a hydrated state in order to have sufficient proton conductivity. Typical of these systems are the perfluorinated sulfonic acid (PFSA) membranes made by DuPont, the standard being Nafion<sup>®</sup>. 3M, Gore, Asahi, and others make membranes with similar properties and characteristics. Intermediate temperature PEMFCs that are capable of operating above the boiling point of water in the region of 120°C are under development primarily for automotive applications. These are still in the development phase [4], and when available will have a significant impact on PEMFC performance for applications other than just automotive. Still higher temperature PEMFCs (HTPEMFCs) are based upon phosphoric acid-doped materials such as polybenzimidazole (PBI). These proton-conducting polymer electrolyte membranes allow fuel cells to operate at temperatures in the range of 160–200°C, thus providing for a high degree of tolerance to certain fuel cell contaminants such as CO, while at the same time offering a source of high-quality heat [5]. PBI-based systems are discussed in detail elsewhere in this encyclopedia, and will be mentioned briefly below in terms of their potential to simplify system design.

For a basic understanding of PEM fuel cells, it is important to appreciate how the properties of the polymer electrolyte dictate the requirements on the fuel cell, stack, and system design. Here just one but critically important example will be discussed: water management. One of the more well-known and basic properties of the PFSA type of polymer electrolytes is that their protonic conductivity is a strong function of the level of hydration. Sulfonated fluoropolymer membranes are based on a polytetrafluoroethylene (PTFE) backbone that is sulfonated by adding a side chain ending in a sulfonic acid group ( $-\text{SO}_3\text{H}$ ) to the PTFE backbone. The resulting macromolecule contains both hydrophobic regions associated with the backbone and hydrophilic regions associated with the sulfonic acid group. Thus, a hydrated PFSA membrane forms a two-phase system consisting of a water-ion phase that is distributed throughout a partially crystallized perfluorinated matrix phase. As the membrane adsorbs water, the first water molecules cause the sulfonated group to dissociate, forming hydronium ( $\text{H}_3\text{O}^+$ ) ions. The water that hydrates the membrane then forms

counterions that are localized on the sulfonated end groups, which act as nucleation sites. As more water is added, the counterion clusters coalesce to form even larger clusters, until a continuous phase is formed with properties that approach those of bulk water. The level of hydration is measured in terms of a parameter  $\lambda$ , which is the equal to the number of absorbed water molecules per sulfonated group. For  $\lambda = 0$ , there is no water, and this anhydrous form of the membrane is uncommon, since complete removal of water requires temperatures near the decomposition of the polymer. Molecular dynamic simulations indicate that the primary hydration shell around the sulfonated group grows to a level of about  $\lambda$  equal to five waters [6]. As more water is added, for  $\lambda > 6$ , the added water is screened by the more strongly bound water of the primary hydration shell, and it can be considered a free phase. Saturation occurs in Nafion<sup>®</sup> at  $\lambda = 14$ . One of the consequences of this strong coupling is that proton transport within the membrane is accompanied by the transport of water. This effect is known as “electroosmotic drag,” and it complicates the required management of the fuel cell by-products: water and heat.

The mechanism and degree of protonic conductivity change as the level of hydration increases. Kreuer [7] has provided an in-depth review of the physics underlying the basic mechanisms of transport in proton conductors. Transport of the proton can occur by two mechanisms: structural diffusion and vehicle diffusion. The structural diffusion mechanism is associated with the proton “hopping” along water molecules (Grotthuss shuttling). Vehicular diffusion is the classical Einstein diffusive motion. The former is viewed as a discrete mechanism, the latter a continuous mechanism. In the nanosized confined hydrophilic spaces within the membrane, both mechanisms are operative, with the diffusion constant of the discrete mechanism being larger and increasing faster than that of the continuous mechanism at higher levels of hydration. At intermediate and low degrees of hydration, the proton mobility is essentially vehicular in nature. What is important here is that the underlying mechanism of transport in proton conducting membranes changes as a function the level of hydration, and it is fundamental to the overall behavior of the fuel cell and therefore the design of the fuel cell system.

Given this fundamental and strong dependence of protonic conductivity with water content, it is necessary to ensure that the polymer electrolyte is safeguarded against localized drying. Indeed, a situation in which the membrane is in a “drying” mode of operation may lead to a catastrophic “death spiral” failure mode, since lower conductivity leads to higher resistances, increased localized temperatures, and therefore an ever-increasing rate of drying, ultimately resulting in membrane failure. In addition, other failure modes, including attack by hydroxyl radicals, are also accelerated at higher temperature and drier membrane conditions, further exasperating the situation.

Hence, it is necessary to maintain the localized level of membrane hydration through “water management,” the *active* management of the moisture content of the membrane. One method is to design the system so that the relative humidity (RH) of the reactant streams coming into the stack is nearly saturated; an approach used by Plug Power and other fuel cell manufacturers. Another approach is to provide for water injection, an approach used by Intelligent Energy. Or, one can employ porous plates to distribute the product water generated during operation, a technique used by United Technologies. Each of these techniques requires careful stack and system design. The design approach taken likewise needs to avoid the situation where too much water can cause localized “flooding” within the cell, which in turn results in fuel starvation and subsequent permanent performance degradation. In most cases, the requirement to manage the polymer electrolyte membrane water content adds considerable system complexity and cost, as well as sources of unreliability.

### Critical Parameters/Critical Specifications

It is clear that in a typical operating fuel cell, the current–voltage characteristics are dependent on the operating temperature, pressure, flow rate, and composition of the reactant gas streams. These are the system-level *critical parameters* around which the fuel cell system must be designed in order to achieve and maintain the desired electrical state of the system. It is also obvious that the performance of the fuel cell is also dependent on the how well the individual *functional*

*requirements* are met with the *specifications* on the component parts. From a systems design perspective, it is perhaps worthwhile to describe how these terms are related. To motivate this discussion, it is important to note that a system is “technically ready” when all the required critical parameters can be simultaneously achieved through the intended interactions of the buildup of parts whose specifications can be met with ordinary manufacturing methods. It is assumed that the critical parameters are chosen to meet the system specifications, while at the same time avoiding all known failure modes. The extent of the range in the critical parameters under which the system can operate over time, without failure, is called the “operating latitude.”

As a simple example, assume that one can write the equations of physics in the linear form

$$F_i = K_{ij}X_j \quad (8)$$

The terms  $F_i$  and  $X_j$  are termed the *critical parameters*, and the  $K_{ij}$  are termed the *critical specifications* that are required to meet the *functional requirements* of the particular component under consideration. The term “critical” is of course an indication that the specification in question is one of the significant few specifications that will be tracked all the way from the voice of the customer down to the factory floor. In a complete PEM fuel cell system, there can be hundreds of specifications that will need to be tracked during the development of the system. Of these, perhaps a few dozen will remain critical even after the product has launched and is in production.

As an example, consider at the conduction of protons across the membrane and for the moment neglect the electroosmotic drag of water with the proton. In this case, Eq. 8 takes the form of Ohm’s law:

$$d\varphi/dz = -[1/\kappa]i \quad (9)$$

Here  $d\varphi/dz$  is the potential gradient across the membrane,  $i$  is the flux of protons, and  $\kappa$  is the protonic conductivity. The system-level critical parameters (pressure, temperature, flow rate, and composition) establish the conditions for determining the local critical parameters ( $d\varphi/dz$ ,  $i$ ), and these are in turn are determined by the functional requirement that the polymer electrolyte has a protonic conductivity,  $\kappa$ . Typical values lie in the range of  $10^{-1}$  to  $10^{-2}$  S/cm.



Notice that critical parameters are generally *not* dimensions, but are rather forces, and fields, affinities, and fluxes. They are the conditions that are required for operation. Specifications, on the other hand, are the items that an engineer can call out on a drawing: dimensions, conductivities (both electrical and thermal), moduli of elasticity, and so forth. Instead of specifying the protonic conductivity  $\kappa$ , which is fixed by the material chemistry, the design engineer will specify the membrane thickness, thereby affecting not only the resistive losses but also rate the diffusion of gaseous species across the membrane.

As discussed above, the protonic conductivity is known to be a strong function of the moisture content, which again will be influenced by the system-level critical parameters of flow rate, temperature, pressure, and composition. Hence, in reality, a *constitutive equation* of the dependence of conductivity on moisture content is needed to support larger PEM fuel cell models which are essential to account for and keep track of all the complex interactions between the critical parameters and the specifications for the components of the MEA and other critical components. An early constitutive equation defining these interactions was described by Springer et al. [8]. These interactions place stringent demands upon the design of the cell, stack, and ultimately the system. Researchers and engineers have developed sophisticated models in an attempt to describe the current–voltage response of single fuel cells to the significant critical parameters through the specifications of the material properties of the cell components. These models, which are remarkably successful in describing the overall performance of a single cell, are macro-homogeneous in the sense that they use local average properties, such as electrode porosity, conductivity, etc., to describe the materials. Constitutive equations of the conductivity dependence on moisture have been developed to support the fuel cell models [9]. There are also excellent reviews of the models that have been developed to elucidate the various aspects of fuel cell behavior under a variety of conditions, as for example those by Weber and Newman [10] and Wang [11]. In addition, there are specialized models that deal with the root cause of specific failures, such as platinum dissolution [12] and carbon corrosion [13] due to high potentials generated during startup and shutdown [14]. All of

these have added considerable insights into how the PEM system runs and how failures are to be avoided. Finally, there are models of the fuel cell system itself. The models predict full system behavior over the range of system-level critical parameters, and are useful in developing the algorithms needed in implementing model predictive control strategies.

Extensive cell/stack/system-level testing is also needed to assess the system's overall performance. In this case, computational fluid dynamics (CFD) is generally used to guide the cell/stack level testing, aiding in the fuel cell system development process. Some of these models have been made commercially available. In particular, models for PEMFCs can be obtained from CD-Adapco [15], Fluent [16], and others. In fact, knowledge of the cell and stack failure modes, and their dependence on the process critical parameters and critical specifications, is fundamental to designing and developing commercially viable fuel cell systems. A more detailed discussion of failure modes will be addressed in what follows. First, however, it is important to understand the current state of the art relative to the component at the heart of the PEM fuel cell, the MEA.

## Recent Advancements

The operation of the fuel cell involves complex interactions at the numerous embedded surfaces within the cell, including the triple phase boundary. These embedded surfaces can mask the ability to probe the detailed nature of these interactions. Today, processes at the nanoscale are a fertile ground for research, enabled by the tremendous advances made in computational nanoscience through the convergence of “nano” and “info” technology. These advancements in computation have allowed theoreticians to explore electrochemical and electromechanical processes at the micro, meso, and nanoscales, with meaningful numbers of atoms, thereby improving the theoretical understanding of the basic processes involved. Furthermore, the rapidly expanding field of nanoscience and nanoengineering is being applied in a variety of fields, to include the development and characterization of nanoscale materials and structures that have direct bearing on energy generation, conversion, and storage.

## Electrocatalysts

Relatively recent advancements have been made in the understanding of heterogeneous ORR catalysis using the combination of molecular and atomic level simulation tools, nanoscale fabrication methods, and nanoscale characterization techniques. For example, the early experimental work of Markovic [17], combined with the important theoretical work of Nørskov [18], Mavrakakis [19], Neurock [20], and others, has now given a deeper understanding not only of the thermodynamics involved in the ORR, and the likely cause of the associated overpotential, but also how variations in the electronic structure determine trends in the catalytic activity of the ORR across the periodic table. This work, coupled with the insights provided by the seminal work of Adzic [21], Stamenkovic [22], and others in the synthesis and performance characterization of Pt alloys and core-shell architectures, has significantly advanced the understanding of the state-of-the-art catalysts for use in both the HOR and ORR fuel cell reactions.

The exact mechanisms underlying the oxygen reduction reaction are still debatable. However, Nørskov [23] and his group have used density functional theory (DFT) to significantly advance the understanding of the thermodynamics of the various possible reaction pathways, and their dependence on surface properties. In combination with detailed DFT calculations, Nørskov and his team were able to provide a detailed description of the free-energy landscape of the electrochemical ORR over Pt as a function of applied bias. In doing so, they found that adsorbed oxygen and hydroxyl species are very stable intermediates at potentials close to equilibrium, and the calculated rate constant for the activated proton/electron transfer to these adsorbed intermediates accounts for the observed kinetics. On this basis, they were able to account for the trends in the ORR rate for a large number of different transition metals. In particular, they were able to construct “volcano” plots of maximum catalytic activity as a function of the oxygen and hydroxyl adsorption energies, describing known trends and the observed effects of alloying. It is clear that a significant amount of progress has been made in the past several years in the basic understanding of the electrocatalytic processes at the nanoscale. An excellent

summary of this and other work has been given in the recent book on fuel cell catalysis by Koper [24].

The high price of platinum has provided the impetus for reducing the platinum loading by improving the kinetics of the ORR reaction. Considerable experimental effort has gone into the research and development of electrodes that incorporate the advantages of new nanomaterials intended to significantly reduce precious metal catalyst loadings, while maintaining high power densities. Chief among these are electrodes that incorporate platinum alloys with transition metals of various types as a means of increasing catalytic activity, thereby allowing a reduction in platinum loading. Currently, Pt<sub>3</sub>Co alloys supported on carbon are achieving roughly a 2× improvement over the state-of-the-art Pt/C electrodes. Intermetallic architectures of Pt and other materials have been fabricated by a variety of techniques, some of which yield so-called core-shell and skeleton/skin structures. Some of these, which involve monolayers of Pt particles on Pd cores supported on carbon, are reported to have achieved very significant (≈20) increases in mass activities.

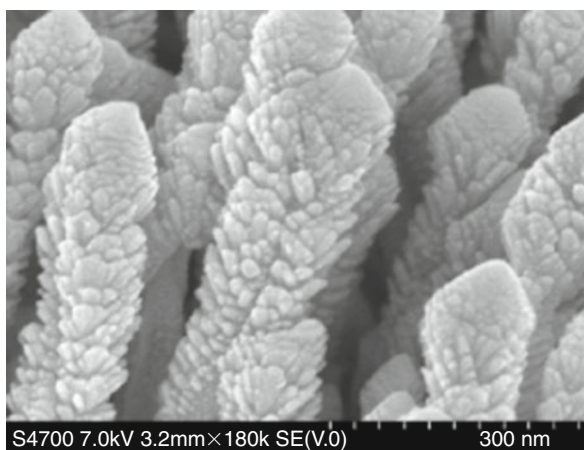
Steady progress has been made experimentally in reducing the platinum loadings, particularly in the hydrogen/air fuel cells used in automotive applications. Gasteiger et al. [25] have discussed the strategies to achieve the nearly fivefold reduction in loading needed in order to achieve automotive cost targets. These strategies, which involve increasing the power density at high voltages by reducing the Pt loading through an increase in the Pt activity, and by reducing mass transport losses in the cell, should also be relevant to backup, motive, and stationary PEMFC applications.

## Nanostructure Electrodes

Most porous electrodes are made up of a mixture of ionomer (for proton conduction), high surface area carbon (for electron conduction), and nanoparticles of catalyst all mixed together to form an ink-like random medium. This ink is then either sprayed or doctor-doped onto either a microporous layer on the gas diffusion media, or applied directly to the polymer electrolyte. In either case, the density of triple phase boundaries is to a large extent left determined by the random nature of the ink and application process.

On the horizon, however, lies a new class of electrodes based upon nanoscale structures that enable both increased performance and better utilization of the precious metals. These so-called nanostructured electrodes are highly structured by design, and enable significant improvement in both performance and durability. To date, the most significant of these structures are those under development by Debe [26] at 3M, wherein the structure consists of a dense forest of crystalline organic whiskers with areal densities on the order of three to about ten billion whiskers per  $\text{cm}^2$ , highly oriented with their long axis normal to the substrate. The whiskers are nanoscale in size (roughly 50 nm by 30 nm by 300–1,000 nm) and provide an interesting support for the Pt catalyst. An image of these whiskers, coated with a platinum catalyst, is shown in Fig. 1 below.

High-resolution TEM studies of the sputter-deposited thin catalyst film on the whiskers indicate that the catalyst films grow as polycrystalline layers that expose highly oriented crystallites. The resulting catalyst “whiskerettes” growing on and at an angle of roughly  $70^\circ$  to the surface of crystalline whiskers represents a very interesting hierarchical nanostructure that is providing significant improvements in performance. As proposed by Debe [27], the source of the activity gain of such a system might have its contribution rooted in the nature of the metal/support interactions.



Polymer Electrolyte (PE) Fuel Cell Systems. Figure 1 3M's Pt-coated NSTF electrode [26]

How well do these nanostructured thin-film (NSTF) electrodes meet the expected performance targets? Current DOE maturity targets for light duty vehicle applications specify performance in terms of platinum group metal loading of  $<0.2 \text{ mg/cm}_{\text{MEA}}^2$ , a durability target of at least 5,000 h (at temperatures  $<80^\circ\text{C}$ ), a mass activity of at least  $0.44 \text{ A/mg}_{\text{Pt}} @ 0.9 \text{ V}_{\text{IRfree}}$ , and a specific activity of  $720 \text{ } \mu\text{A/cm}_{\text{Pt}}^2 @ 0.9 \text{ V}_{\text{IRfree}}$ . At the most recent DOE Annual Merit Review, 3M reported that they had achieved  $0.19 \text{ mg}_{\text{Pt}}/\text{cm}_{\text{MEA}}^2$  loadings on  $400 \text{ cm}^2$  in OEM short stacks, 5,000 h of durability (20  $\mu\text{m}$ , 850 EW with no stabilizers),  $0.24 \text{ A/mg}_{\text{Pt}}$  with PtCoMn alloy ( $0.40 \text{ A/mg}_{\text{Pt}}$  with a new  $\text{Pt}_3\text{Ni}_7$  alloy that needs further development), and specific activity of  $2,100 \text{ } \mu\text{A/cm}_{\text{Pt}}^2$ . This is a substantial improvement in performance over the conventional Pt/C paradigm that has dominated fuel cell MEA designs for the past 15 years. There are water management issues with the thin electrode structure, as the reduced thickness of the electrode results in flooding more easily at cold, humid conditions while at high current density. 3M is studying a unique set of operating conditions and GDL properties on the anode side in order to reduce the impedance of water transport from the cathode to the anode, in order to improve the operating latitude with respect to cathode flooding [28].

It is important to note that in the pursuit of other types of highly ordered nanostructures, researchers are also taking aim on the issue of the catalyst support from the perspective of its stability. As will be discussed below, the usual carbon materials used to support the catalyst nanoparticles corrode under certain operating conditions associated with startup and shutdown of the fuel cell, or conditions of localized fuel starvation, and so it is desirable to have a support which is more stable under these dynamic situations. Accordingly, researchers are developing nanostructures in which the typical carbon material is replaced with materials with greater stability. Included in this set of support materials are the use of carbon nanotubes [29] and ceramic materials such as  $\text{TiO}_2$  [30]. It is anticipated that advanced nanostructured electrodes will try to emulate the excellent results and the lessons being learned in the case of the NSTF work underway at 3M, while avoiding problems of water management.

### Low and Intermediate Temperature Proton Conducting Polymer Electrolyte Membranes

There is also significant research underway addressing the development of alternative classes of polymer electrolytes for fuel cells. For example, sulfonated aromatic polymers, sulfonated polyimides, proton conducting membranes carrying phosphonic acid groups, and polyphosphazene polymer electrolytes are all being examined as lower cost, more durable alternatives to PFSA-based membranes. There are excellent reviews of the progress being made in the development of these acid-based polymer electrolytes [31].

As noted, PFSA-based polymer electrolyte membranes used in current PEMFCs require thermal and water management systems to control temperature and keep the membrane humidified. These extra components increase the weight and volume of the fuel cell system and add complexity. The cost and complexity of the thermal and water management systems could be minimized if the fuel cell operated at higher temperatures (up to 120°C) and at lower relative humidity (RH). The US Department of Energy has therefore initiated a major effort to develop new membranes that can operate at temperatures up to 120°C without the need for humidification. While it is desired that the fuel cell membrane operate without external humidification, it is recognized that the water generated by the fuel cell itself can be utilized to provide some humidification of the membranes. It is expected that under proper operation, recycling product water from the cathode to the inlet air can provide roughly 25% RH inside a stack operating at 120°C with inlet feeds at a water vapor partial pressure of  $<1.5 \text{ kPa}$ . This reduces the burden for performance with very low water content, but operation at this low RH still remains a major challenge.

A variety of strategies are being pursued. It is known from extensive molecular level modeling of Nafion<sup>®</sup> that membrane microstructure has a substantial effect on conductivity. With block copolymers, McGrath [32] has observed that as the block length increases, the performance under partially hydrated conditions increases, suggesting the presence of hydrophilic domains at higher block lengths through which protons can be transported along the sulfonic acid groups and water molecules. Utilizing block copolymers

consisting of hydrophilic oligomers and hydrophobic perfluorinated oligomers, McGrath and his coworkers have prepared a block copolymer that exhibited higher proton conductivity than Nafion<sup>®</sup> 112 at 80°C at all RH values (30–95%).

FuelCell Energy is pursuing a rather sophisticated approach utilizing a four-component composite consisting of a copolymer, a support polymer, a water retention additive, and a protonic conductivity enhancer. The copolymer, which is intended to provide the basic building block for the membrane, is an advanced perfluorosulfonic acid with significantly higher conductivity than state-of-the-art polymers. The support polymer is intended to give a stable cluster structure and to enhance mechanical properties. The functionalized additives are designed to retain water at low RH conditions and to enhance the composite membrane's proton conductivity by providing an alternate proton conduction path for the efficient transport of proton at high temperature as well as at subfreezing conditions. Progress to date has led to a membrane with conductivity about a factor of 2 better than Nafion<sup>®</sup> 112 at 120°C [33].

All of these approaches still rely on water for conduction. Systems utilizing phosphonic acids, heteropolyacids, protic ionic liquids, and heterocyclic bases do not rely on water for conduction. One of the major issues for these systems, however, is that the acid or base aiding proton transport is generally water-soluble. The acid or base must be immobilized for use in transportation applications where condensation of liquid water under some of the operating conditions is inevitable. However, enough mobility must be retained by the active group to be able to participate in proton conduction.

In addition to these more advanced studies, progress has been made in developing improvements to the current PFSA polymer electrolyte membranes currently employed in PEMFCs, particularly with regard to degradation caused by peroxide generated either electrochemically or chemically as the result of oxygen crossover to the anode. It is generally accepted that the membrane degradation is the result of the subsequent formation of hydroxyl radicals that attack the side chains of the polymer [34–36]. Here again, dry conditions greatly accelerate the rate of degradation.

What is not generally agreed upon is the location and origin of the hydrogen peroxide formation. Two theories are dominant. One is that the peroxide is formed mainly at the anode due to the diffusion of oxygen across the membrane, at which point the peroxide can form either chemically or electrochemically by the reaction of oxygen with hydrogen. This hypothesis is supported by the fact that the peroxide yield is greatly enhanced in the anode potential region ( $<0.2$  V) of the ORR, as evidenced by rotating ring disc electrode experiments [37]. The other is that the peroxide is formed at the cathode. Experimental results from several sources support this hypothesis. Liu et al. [38], Mittal et al. [39], and Miyake et al. [40] have commonly reported that the membrane degradation rate is higher for an MEA catalyzed only at the cathode side than for one catalyzed only at the anode side. This fact seems inconsistent with the hypothesis that peroxide is formed at the anode. Yu et al. [41], in an effort to identify the origin of the radicals which decompose the membrane, performed a series of experiments after which they concluded, based upon energy dispersive X-ray analysis and IR spectroscopy, that the degradation begins on the cathode and progresses inward. They also found that the formation of radicals within the membrane or near the anode was very low or absent. They postulated the formation of undesirable free radicals involving the catalyst, similar to the hypothesis of Liu and Mittal.

Any model of the degradation process would need to propose that formation of hydrogen peroxide forms by distinct mechanisms in the cathode and anode. The commonly accepted mechanism is that the peroxide then forms radicals through Fenton reactions involving metal-ion impurities. The radicals then participate in the decomposition of reactive end groups in the membrane, to form, among other species, hydrogen fluoride, which can be detected in the product water. Higher fluoride release rates correlate with higher rates of membrane degradation. The degradation occurs through the “unzipping” of the polymer backbone and the cleavage of the polymer side chains. The two conditions that accelerate this degradation mechanism are dry conditions and high temperatures, thinning the membrane until “pin holes” form, allowing gas crossover and cell failure. Therefore, as emphasized earlier, PFSA-like polymer electrolyte materials must be kept under good temperature control and maintained hydrated.

The susceptibility to peroxide radical attack has been attributed to a trace amount of polymer end groups with residual hydrogen-containing terminal bonds [42]. It is at these sites that decomposition is initiated. DuPont has reported that the number of hydrogen-containing end groups can be reduced by treating Nafion<sup>®</sup> with fluorine gas. 3M has been improving their PEMs by modifying the end groups in the membrane. 3M is investigating modified polymer structures to try to control membrane morphology and has developed a new ionomer with a shorter side chain than standard perfluorosulfonic acid (PFSA) membrane ionomers without the pendant  $\text{CF}_3$  group. They have reported that this structure provides a higher degree of crystallinity and allows for lower equivalent weight membranes with improved mechanical properties and durability under hot, dry conditions. Solvay Solexis also has a short side-chain PFSA membrane, Aquivion<sup>™</sup>, that resists peroxide radical attack. In addition, companies are not only modifying the chemical structure but are also incorporating proprietary stabilizers to mitigate the effect of peroxide and the attack by its radicals.

### High-Temperature Polymer Electrolyte Membranes

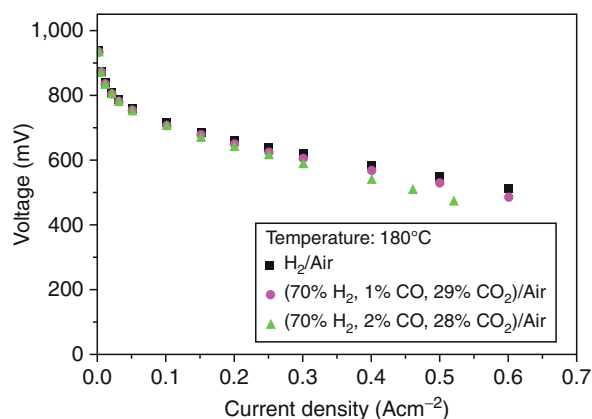
Early work on the properties of proton conducting acid polymer blends was carried out by Lassegues et al. [43, 44]. The first detailed study of the effect of doping a high-temperature polymer membrane such as polybenzimidazole (PBI) with a phosphoric acid in order to achieve proton conductivity in a range suitable for use as a fuel cell was conducted at Case Western Reserve [45]. Polymer electrolyte membranes that do not rely on water as the basis of their proton conductivity offer the potential advantage of simplifying the fuel cell system, if these systems can be operated without complex external humidification schemes. Furthermore, because these systems run at high temperature, they can tolerate higher levels of CO in the reformat, thus further simplifying the system in terms of reforming and control. In addition, the higher operating temperature simplifies thermal management issues and provides a source of high-quality heat for combined heat and power applications. Consequently, BASF [46] and other companies are developing PBI-based proton conducting membranes for

high-temperature stationary fuel cell applications. Several companies (for example, Plug Power, ClearEdge) are developing products based upon these materials.

New high-temperature polymer electrolyte membranes based on aromatic polyether polymers and copolymers containing polar pyridine moieties in the main chain have been developed by Advent Technologies for use in HTPEM fuel cells. These are based on the idea of creating acid–base interactions in order to obtain high proton conductivity. The newly developed polymer products are called Advent TPS<sup>®</sup>. Such materials combine the excellent film-forming properties with high mechanical, thermal, and oxidative stability and the ability to be doped with phosphoric acid. Highly conducting polymer electrolyte membranes were produced after treatment with phosphoric acid amounts that could be controlled by varying the pyridine-based monomer content. The polar pyridine groups strongly retain the phosphoric acid molecules, due to their protonation, thus inhibiting leaching out of the phosphoric acid.

The performance based on Advent TPS<sup>®</sup> MEA operating at 180°C with pure hydrogen or reformat with different CO contents and air feed gases is shown in Fig. 2. Here again, because of the high operating temperature, this system tolerates up to 2 vol% CO poisoning without a significant decrease of performance.

As is well known, and as discussed by Neyerlin and others, there is an increase in polarization losses with



**Polymer Electrolyte (PE) Fuel Cell Systems. Figure 2** Polarization curves of Advent TPS<sup>®</sup> at 180° with H<sub>2</sub> or reformat gas and air, under ambient pressure ( $\lambda_{\text{H}_2} = 1.2$ ,  $\lambda_{\text{air}} = 2$ )

phosphoric acid-based fuel cells, in comparison to low-temperature PFSA-based fuel cells, and this effect is thought to be due to the presence of phosphoric acid and/or its anions that adsorb onto the surface of the catalyst [47]. Because of this, high-temperature stacks based upon phosphoric acid–doped polymer electrolyte membranes are larger in order to get the same power output. The overall question is whether the benefit in system simplification overcomes the need for larger stacks, so that there is an overall net system benefit. Reducing the effect of the adsorbed anion species would, of course, have significant benefit at the stack and system levels.

### Alkaline Polymer Electrolyte Membranes

There are also efforts underway to develop alkaline-based polymer electrolytes, which enables the replacement of precious metal catalysts by catalysts based on highly inexpensive metals, such as Co, Fe, and Ni. In particular, Celler Technologies (Caesaria, Israel) is pursuing the development of a new alkaline technology by transitioning from the proton conducting polymer electrolyte membrane and ionomer to an OH<sup>-</sup> ion conducting membrane and ionomer. The alkaline ion-conducting polymer electrolyte membrane provides a much more benign chemical environment which lowers the risk of instability caused by the corrosion processes prevalent in the highly acidic environment of the PEM fuel cell.

Alkaline-based polymer electrolyte membranes are relative newcomers on the scene. It has been widely believed that the quaternary ammonium hydroxide functional group (RN<sub>4</sub><sup>+</sup>, OH<sup>-</sup>), which is the one used in most anion exchange membranes, is “self-destructive,” because the OH<sup>-</sup> ion is likely to attack the RN<sub>4</sub><sup>+</sup> cation. In addition, the specific conductivity of an OH<sup>-</sup> conducting ionomer was suspected of being at least a factor 3–4 lower than that of the H<sup>+</sup> conducting ionomer, thereby setting a limit on power output. It was also suggested that, as the AEMs developed to date have been based on hydrocarbon, rather than fluorocarbon backbones, the preparation of effective and stable membrane/electrode assemblies would present a significantly tougher challenge versus the case of PFSA ionomers. Historically, there has always been a concern having to do with the effect of CO<sub>2</sub> from

the air feed on the OH<sup>-</sup> conducting ionomer. Since conversion of the OH<sup>-</sup> ion in the alkaline ionomer to bicarbonate (and/or carbonate) ion is a very likely process, cell performance could suffer from the effects on both lower ionomer conductivity and, particularly, the kinetics of electrode processes.

Tokuyama, a Japanese company specializing for a long time in membrane technology for electrodialysis and desalination, has undertaken development of AEMs in OH<sup>-</sup> form, targeting fuel cell applications. Los Alamos National Laboratory researchers have shown that proper selection of the functional group for the AEM could secure stability of the ionomer, maintaining the functional group population under demanding chemical conditions. Cellera has so far proved that factors limiting AEM-FC longevity do not include the chemical stability of the alkaline membrane. Furthermore, the membrane resistivity is not more than two times higher versus Nafion, which is an acceptable level for fuel cell development. Finally, it has apparently been shown that the effects of CO<sub>2</sub> from air on the performance of the AEM-FC can be mitigated by using a combination of tools upstream the stack and in the stack. This is the issue that has plagued most alkaline-based systems which get one of their reactants from the air, including metal/air batteries. The approach developed by Cellera for handling CO<sub>2</sub> upstream of the stack supposedly totally avoids any need of routine replacement of system components. Cellera has so far demonstrated power densities of 200 mW/cm<sup>2</sup> at a cell voltage of 0.4 V, with no platinum catalysts for a hydrogen/air stack, with no addition of liquid electrolyte, and with no humidification from an external source [48].

### Fuel Cell System Design

The *process* of designing of a fuel cell system follows the general design process for any commercial product in which a phase-gated, product development process (PDP) guides activities from initial concept to launch. The PDP is one of the most interesting and well-studied business processes, and usually ends up being tailored to each company's appetite for speed and time to market, balanced by its desire to mitigate risk. Product development is very much a social science, one that integrates business, technology, engineering, and

human behavior [49]. Many attempts have been made to study its characteristics and improve the manner by which products are brought to market [50]. Even today, it remains a subject of research in most business schools. Before these steps are undertaken, however, usually a *product strategy* is developed which aligns to both the business strategy and the business model of the organization. This product strategy generally takes the form of a *platform strategy*, which shows the relationship between multiple products expected to be developed and launched over a 5–10 years time horizon [51].

### System Requirements and Architecture

The design of the fuel cell system really begins with the “Voice of the Customer” (VOC). For fuel cell systems, there can be many potential customers: the end user, the OEM, or an intermediate service provider. In any case, by the voice of the customer is meant that quality expectation taken directly from the customer, properly evaluated, and deployed within the product development process. One such technique for doing this is the well-known Quality Function Deployment (QFD) process which, if properly used, enables the VOC to be deployed all the way to the factory floor. Without knowing and agreeing on the key customer requirements upfront, the fuel cell system design process cannot be successfully completed.

A very much abbreviated and simple example of one of the first steps in the process of developing fuel cell system specifications is shown in Fig. 3. This example is for a residential stationary fuel cell application, but similar requirements are relevant in most other fuel cell applications. Here is shown the high-level VOC, gathered by some technique such as the use of customer interviews or focus groups and the system attributes that will influence their achievement, arranged in order of importance. As shown, some of the key system attributes that are critical to quality (CTQs) and that influence the customer's demand for clean, quiet, safe, and affordable energy are related to meantime between failure, major component life, load-following capability, turndown ratio, operating and maintenance cost, efficiency, startup time, audible noise, and agency compliance. Of course, there are a host of other attributes which need to be specified and met, including the expected ranges in ambient temperature and operating

Developed country end user CTQs	Customer weight															
		System reliability	FC control and load control	Service provider program	System serviceability	System availability	System design	System kW rating	Cost - capital and installation	System packaging	Cost - O&M	Certification	System efficiency	Customer service	Cost - fuel	System installation
Simple to operate	5	9	9	9	3	9	3	3	3	3			3			
Safety	5	9	9	9	9	3	9	3	3	9	3	9	1	1	3	
Monthly bill <= current utility bill	5	9	9	1	1	3	1	9	9	3	9	9	1	9	1	
Supply 100% of electricity needs	5	9	9	9	3	9	3	9	1				3			
Responsive service	4	3	3	9	9	3	3		3	3			3			
Better availability than current alternative energy source	3	9	9	9	9	9		3	3	1	1		3		3	
Ascetics; appearance and physical size	3	3	3		9	3	9	9	9		9	3			3	
Environmentally friendly	2	1				3	1				1	9				
Totals		230	228	203	170	159	134	134	116	108	90	77	77	61	45	38
rank		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

### Polymer Electrolyte (PE) Fuel Cell Systems. Figure 3

Example of a limited set of attributes for a residential fuel cell system

altitude, expected number of starts and stops, and the quality of air and water needed for operation. All of these external requirements drive the selection of a technology set which, when optimized, needs to provide for acceptable operating latitude over the life of the system.

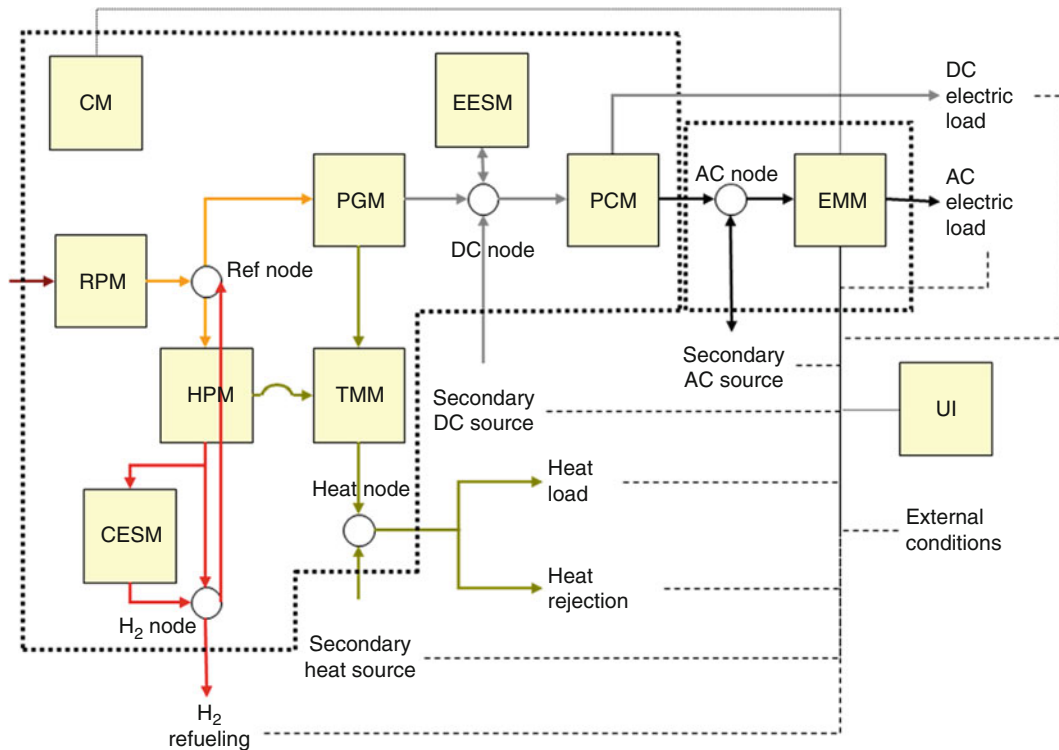
One of the key requirements for a successful platform strategy is a well thought-out platform *architecture*. The platform architecture enables that product strategy to be executed. As an example, consider the high-level architecture shown in Fig. 4 below. Here, one sees the functional breakdown of the system into the major components, which in turn defines requirements for each module in the system.

In this figure is shown several major subsystems typical of a PEM system. Included are the reactant processing module (RPM), which determines the pressure, temperature, flow rate, and composition of the reactant streams which are delivered to the power generation module or PGM. Included in the RPM are the desulfurization, reformer and shift reactors, and the required number of heat exchangers which are needed to manage the reforming process. More will be said of this below. The PGM will include the stack and its manifold, and any other subsystems needed to ensure

the stack can function properly, such as the subsystems that provide for cathode recirculation and water management. Also shown in this figure is a hydrogen pump module (HPM) that can be used to purify and compress hydrogen for storage in the chemical energy storage module (CESM) for refueling a hydrogen fuel cell car, and a thermal management module (TMM) that provides the methods of capturing and directing the flow of heat to either a heat load (such as the RPM) or heat sink. Electrical energy generated by the fuel cell can either be stored in the electrical energy storage module, EESM, or passed through to the power conditioning module (PCM) where it is converted to high-quality power for alternating current loads. A control module (CM) ensures that the required electrical state of the system is achieved over time, as defined by the energy management module (EMM). The CM provides the proper control of the pressure, temperature, flow rate, and composition of the fuel and air streams going into and out of the system. The result of all of this is made visible to “customer” through an appropriate user interface module (UI).

This is a high-level architecture. If the detailed lower-level subsystem requirements that support this “meta system” are well thought out, it is possible to





**Polymer Electrolyte (PE) Fuel Cell Systems. Figure 4**  
A high-level fuel cell system architecture [62]

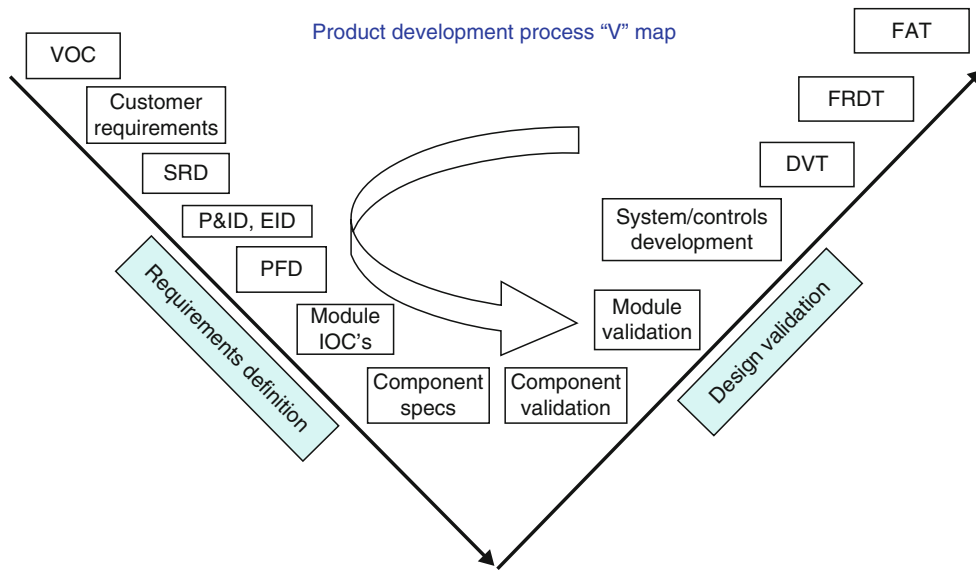
develop *multiple products* off of the same set of hardware and software components, thereby minimizing the product acquisition spending (PAS). For example, at least four products can be derived from this architecture, including a home refueling system, a residential CHP system (eliminating CESM and HPM), a hydrogen fuel cell system for backup applications (further eliminating the RPM system), and a stand-alone hydrogen generator (retaining only the RPM, TMM). Of course, in each case, the controls and user interface have to be modified for the intended application. Furthermore, system-level specifications need to be compatible over the ambient temperature range, etc.

Where the intended range of applications does not allow for a platform architecture at the meta-system level, the platform products then need to be defined at the system level. In fact, the concept of a platform-based architecture can be pursued at the component as well as the system level. Ballard, for example, defines its platforms at the stack level. It is currently advertising

five stack platforms. Each stack platform has a different voltage range, reactant pressure (air and fuel), and cell aspect ratio, optimized for the intended application, whether it be for backup (telecommunications), motive power (forklift trucks), transportation (buses), or residential cogeneration.

While the overall architecture defines the interrelationships among form, fit, and function, *systems engineering* supports the development and achievement of the product specifications. The systems engineering required to enable a product platform strategy is generally quite extensive, requiring many iterations using simulations to guide the final design at the system, subsystem, and component levels. In general, the requirements for a fuel cell system are quite extensive and complex, and these need to be defined, allocated, and enforced all through the development process.

The key steps in the product development process are shown below in Fig. 5. Most activities are focused on two types: the systems engineering associated with requirements definition, ending with components



**Polymer Electrolyte (PE) Fuel Cell Systems. Figure 5**  
Product Development Process Map [62]

specifications, and on the actual design and verification process, ending with a successful in-house design verification test (DVT), a field readiness demonstration test (FRDT), and a factory acceptance test (FAT) to validate launch readiness. A key step in the overall process is the development of the systems requirements document (SRD), which enables the development of the process and instrumentation diagram (P&ID), and its companion, the process flow diagram (PFD). The PFD, which is the output of the process engineering analysis using computational tools such as ASPEN<sup>®</sup>, captures the concepts in flow sheet format and summarizes the heat and material balances characterizing the system design. In short, the PFD, which is the link between concept and reality, defines the critical parameters at each step in the process.

Early on in the process, it is useful to organize the systems engineering effort in pursuit of a collection of interlocking *functionally important topics* which are system-level attributes which need to be allocated to subsystems within the system. One example would be the allocation of efficiency targets between the various major modules. Another high-level example of such an allocation would be the “flow down” of the requirements for cost and reliability. Here, the percentage of total allowed cost and failure rate is allocated among the major modules, based upon an initial test of

reasonableness followed by refinement during the development process as further information becomes available. Clearly, the initial allocation of the functionally important topics is important, because it drives engineering behavior and subsequent design modifications and eventually, time to market. The system may have other high-level requirements, such as size, weight, noise, and appearance, and these early requirements are met through systems engineering efforts. Management of the allocation process is achieved through the use of artifacts such as *input/output/constraint* charts (IOCs) that keep track of both functional (signal) and dysfunctional (noise) attributes of each subsystem, subject to the constraints imposed by the overall system. They track the interactions between subsystems in terms of the critical parameters, and are used to define and guide the robust design optimization process, eventually becoming the contract between the subsystem team and the system engineers. Again, a combination of careful testing and simulation is required to ensure convergence to the subsystem and system-level allocations.

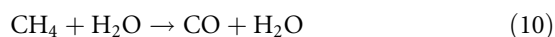
### Reforming Considerations

In many cases, hydrogen in its pure state is not normally available, in which case it must be obtained by the

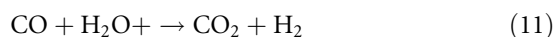
reforming process of stripping hydrogen off a hydrocarbon fuel such as methane, propane, ethanol, or liquid petroleum gases (LPG) or kerosene to extract the hydrogen. Generally, there exist sulfur species that are naturally occurring in the fuel (COS, H<sub>2</sub>S), or are added as odorants (mercaptans, tetra-hydro-thiophene, THT). Since sulfur poisons the catalysts in the system, the first step in the reforming process generally involves the removal of harmful sulfur-containing species. The level to which sulfur needs to be removed depends upon the operating temperature, but in general it is safe to assume that for PEM systems the level needs to be less than about 50 parts per billion (ppb) [52]. To achieve this level of sulfur impurity, there are several available technologies. At ambient temperatures, these include physical adsorption beds using activated carbon and zeolites, and chemisorption beds incorporating nickel-, copper-, and iron-based sorbents. At higher temperatures, one alternative is to use hydrodesulfurization with Co-, Mo-, and Ni-based catalysts. Another option is thermal swing absorption. Ideally, it is desirable to have a single technology in order to have a fuel flexible system that enables the fuel cell system to be sited in a variety of geographical locations. Seasonal and regional variations in sulfur content, as well as catalyst material cost and toxicity, make this a difficult engineering problem. Each of these technologies is well known, and each has its strengths and weaknesses. All add considerable capital cost and complexity to the system, not to mention annual maintenance cost which can also be a significant expense. Consequently, there is a need for continued research into the discovery and development of effective, low-cost desulfurization for fuel cell applications requiring them.

Assuming that the fuel has been adequately desulfurized, there are essentially three alternative processes for extracting hydrogen: steam reforming (SR), catalytic partial oxidation (CPO), and autothermal reforming (ATR). SR is an endothermic process. Heat must be supplied to the reactor, and this is normally accomplished by combusting an additional amount of fuel in a separate, but thermally integrated reactor. Usually this reactor burns the unused hydrogen which flows through the anode chamber of the fuel cell stack. This component is referred to as the “anode tail gas oxidizer,” or ATO. The ATO is generally considered an integral part of the reactant processing module.

The steam reformation reaction is reversible, and the product gas is a mixture of hydrogen, carbon monoxide, carbon dioxide, and water vapor. In the case of methane, the steam reforming step is



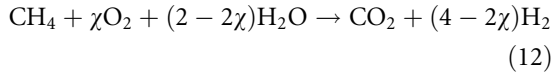
The amount of CO in reformat is a system-level critical parameter, as CO interferes with the hydrogen oxidation reaction by blocking reaction sites on the platinum catalyst surface. To remove the high levels of CO, this reformat stream is then fed to a “shift” reactor which reduces the CO concentration and makes additional hydrogen:



Following the “shift” reaction, the reformat is generally passed through a preferential oxidation (PROX) step to further reduce the CO content to an acceptable level so as not to poison the anode catalyst. For low-temperature PEM systems, this level is typically less than 10 ppm. To increase the tolerance to the residual CO, ruthenium is added to the anode catalyst layer, but even this is not sufficient to deal with the amount of CO in the reactant stream. In the case of low-temperature PEM systems, there is generally a small amount of “air bleed” fed into the reactant stream to oxidize any residual levels of CO that adsorb onto the catalyst. It is desirable to keep the air bleed to a minimum to reduce the formation of peroxide. Pulsed air bleed techniques have been studied to achieve this purpose [53]. Depending on the reformer design, there may be a high temperature as well as a low-temperature shift, as well as a single- or two-stage PROX.

Unlike steam reforming, CPO is an exothermic process. It is essentially a combustion, but with a less than stoichiometric amount of oxygen. As in SR, the reformat gas must go through a shift reaction to produce more hydrogen, though a CPO produces less hydrogen than SR, and unlike SR, the product gas contains relatively large amounts of nitrogen from the air used in the process.

Because CPO is an exothermic process, and SR is an endothermic process, these two can be combined in the “autothermal” reforming (ATR) process. The overall reaction of an autothermal process for methane, with the shift reaction included, is



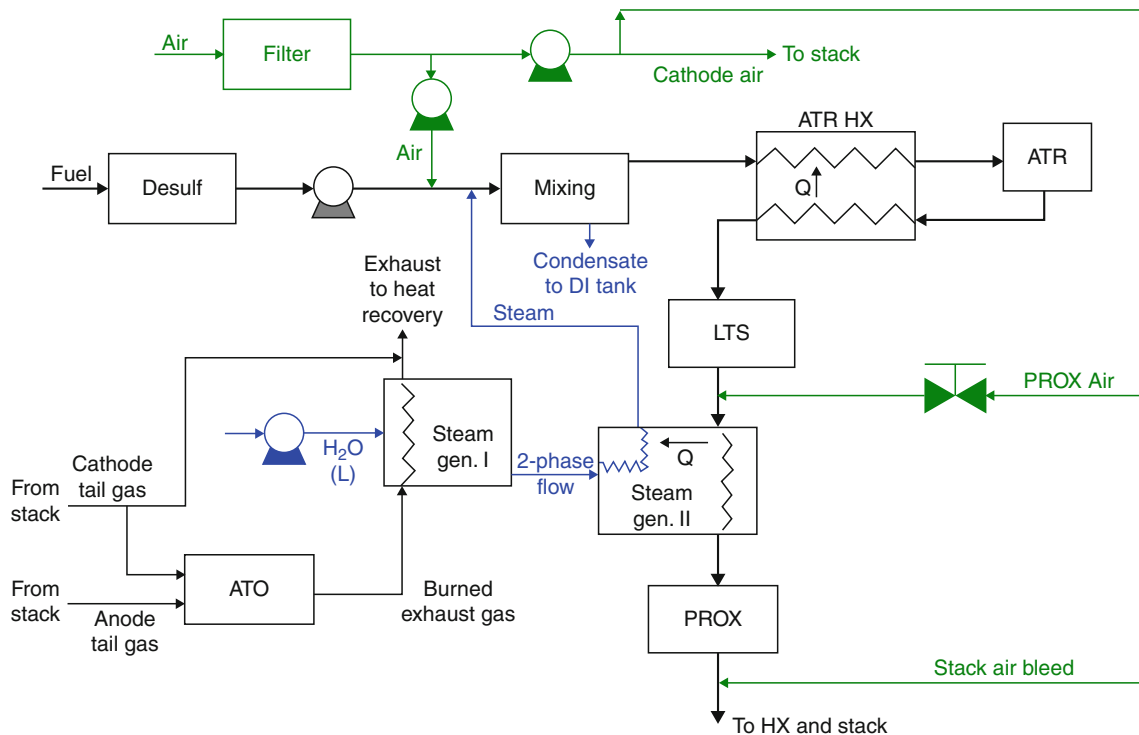
In this equation,  $\chi$  is the number of moles of oxygen per mole of fuel. Its value determines if the reaction (12) is exothermic, endothermic, or thermoneutral. In practice, the value is chosen to avoid the risk of carbon formation in the reactors, and the amount of steam added is usually in excess of that theoretically required. All of these process steps use specially designed catalysts to facilitate the reaction steps. Catalyst designs are in many cases proprietary, although in most cases references to catalyst materials used in similar reactions can be found in the literature.

For many applications requiring the need to reform a logistical fuel, the reactant processing module represents a significant percentage of the total system cost. This is because the reactant processing module must provide reformat with the proper level of humidity over the entire load range, the required level of CO within the constraints of a single stage partial oxidation (PROX) step, have rapid startup and load-following

capability with near zero emissions for  $\text{NO}_x$ , and this all without resulting in coke formation. To meet the efficiency requirements of the overall system, the reactant processing module must also be thermally integrated into the rest of the fuel cell system. This is a very demanding set of requirements, and the cost of the reactant processing module typically ends up being about one-third of the total system cost.

It is well known that most Japanese residential  $\mu\text{CHP}$  fuel cell systems use steam reformers, principally due to their high efficiency. Adachi et al. [54] (2009) have recently published the results of their effort to design an autothermal reforming fuel processor for a 1 kW  $\mu\text{CHP}$  residential system. They designed and bench tested an ATR system capable of achieving 80% efficiency (on a higher heating value basis) and a reformat containing 48% hydrogen (dry basis) and less than 5 ppm CO. They have studied the ATR principally to address the issues around startup time and the daily start/stop requirements.

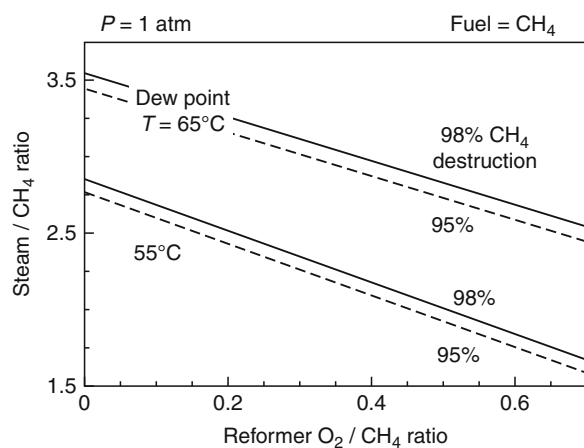
A simplified schematic of an ATR-based reformer system is shown below in Fig. 6. Shown in this figure



**Polymer Electrolyte (PE) Fuel Cell Systems. Figure 6**  
Simplified schematic of an ATR-based reformer subsystem [62]

are the desulfurization subsystems, the ATR reactor, a low-temperature shift (LTS) reactor, PROX, ATO, low-pressure steam generators to meet the reformat due point requirements, and the appropriate heat exchanges.

In practice, there are several critical operating requirements placed on the reformer system. First, the reformat must contain a negligible amount of carbon monoxide (and oxygen) to avoid poisoning the anode catalyst. Likewise, the fractional conversion of fuel in the reformat needs to be specified in order to minimize unburned hydrocarbon emissions. From a system control perspective, it is also desirable to deliver a predictable amount of hydrogen for a given fuel input. Finally, because of the dependence of the PEM on the level of hydration, the reformat dew point must also be controlled to a specified value. Given these specifications, Feitelberg and Rohr [55] have shown that two variables, the steam to carbon and oxygen to carbon ratios, define a unique “operating line” for the fuel processor. The operating line defines the required relationship between steam-to-carbon and oxygen-to-carbon ratio for a reformer that meets fuel conversion and reformat dew point specifications. An example of the operating line [56] for several dew points is shown in Fig. 7. Operation above this line means that the dew point is greater than that specified, implying that energy has been wasted vaporizing more water than is needed. Operation below the operating line means that the reformat dew point is lower than specified,



**Polymer Electrolyte (PE) Fuel Cell Systems. Figure 7**  
Reformer operating line analysis [52]

implying that one should expect reduced MEA life or premature failure. When the PEM system is operating on the line, and the heat generated in the ATO just balances the heat consumed in the reformer, the reforming process has achieved the maximum theoretical efficiency.

In this regard, their analysis has shown that the maximum theoretical efficiency of a SR-based system is only about one percentage point higher than the maximum theoretical efficiency of a typical ATR-based fuel processor. Hence, the ATR technology seems to be a good compromise in performance and efficiency.

### Technical Readiness

A key milestone in the development of a fuel cell (or any other) system is the achievement of technical readiness; that point in the development cycle when commitments to volume manufacturing can be made with confidence. For this purpose, it is meaningful to define this state of knowledge as:

- ▶ Technical readiness is achieved when all critical failure modes are known, and when the critical parameters that are required to avoid those failure modes over the required operating range can be achieved through the intended interaction of specified parts which can be manufactured with acceptable process latitude.

In discussing PEM fuel cells for stationary, transportation, and portable applications, it is worthwhile to do so in terms of defining those conditions needed to not only meet the intended application, but also to avoid the critical failure modes.

### Freezing

One of the customer-driven requirements associated with fuel cells intended for stationary and transportation applications is the ability to not only survive subfreezing temperatures, but to gracefully startup from subfreezing conditions and operate over a large number of expected freezing/thaw cycles. Inasmuch as water is the product of the fuel cell reaction, it is normally carried into the stack with humidified reactants, and is fundamental to the transport of protons in the membrane. Most of the work on defining the operating conditions needed for surviving freezing conditions has focused on the behavior of the water within

the principle stack components: manifold, plates, flow fields, and MEA. There are three types of water in PFSA-type polymer electrolyte membranes. Water that has strong interactions with the ionic groups in the polymer can withstand temperatures well below the normal freezing point of water without freezing. Water that finds itself within the nanoscale channels of the polymer electrolyte has a freezing temperature that is dependent on the size and nature of the channel, which in turn is dependent on the degree of hydration of the polymer. This water will therefore have a range of temperatures over which it will freeze, with some of the water having its freezing temperature depressed by the presence of the hydrated ions. Finally, there is “free” water that behaves like bulk water. The result is that there is a distribution of freezing points within the PEM, with no specific “freezing point.” The PEM proton conductivity, for example, varies continuously over a wide range of temperatures, depending on the level of hydration. The proton conductivity below freezing can be an order of magnitude lower than that for normal operation [57], indicating that a large fraction of the water in the membrane is frozen. However, even in this state, the polymer electrolyte membrane is capable of carrying current, increasing with increased thawing.

Next to the membrane itself, the most critical components in the fuel cell stack relative to startup from freezing and cool down to subfreezing temperatures are the catalyst layer and the GDL. As previously discussed, the GDL is one of the few critical components in the system that can be designed to affect proper water management, including frozen water that exists during storage or operation from subfreezing conditions. However, only a few studies have been published on the effect of freeze thaw cycles on the properties of the GDL and its subsequent effect on the durability of a PEM fuel cell during cold startup [58, 59]. In fact, due to the wide range of operating conditions and the difference in fuel cell design approaches, reported impacts of freeze thaw cycles on MEA performance vary considerably.

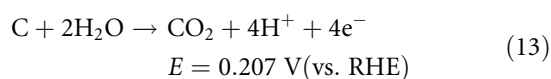
One of the most descriptive accounts of the factors influencing the cold start behavior of low-temperature PEM fuel cells has been given by Mao and Wang [60]. They develop a lumped analysis model of the cold startup of a PEM fuel from subfreezing conditions. The model accounts for the rate of water production, removal and conversion during startup. The condition

for a successful startup is to have the fuel cell raise its temperature beyond the freezing point before being shut down by the formation of ice in the catalyst layer. By accounting for the heat and mass balances within the catalyst layer, the authors are able to investigate the conditions under which ice forms in the catalyst layer, thus restricting the flow of oxygen to the reactions sites. In effect, the authors indicate how to calculate the critical parameter requirements for a successful startup from subfreezing conditions. Initial membrane moisture content, heating rate, and thermal masses need to be chosen so that the temperature within the catalyst layer can reach the melting point before ice can fill up the pores within the catalyst layer and the GDL. Clearly, these critical parameters will depend upon the critical specifications of the components within the MEA, as well as the thermal properties of the bipolar plates.

There are a variety of system-level strategies for dealing with freezing, including purging of the stack at shutdown to remove liquid water, heating the stack prior to or during warmup with auxiliary energy sources, preventing freezing conditions from occurring (standby, idling, etc.) Each of these strategies, however, results in a decrease in overall system efficiency.

### Start/Stops

Every fuel cell system needs to be started up and shut down in a way that meets the requirements of the customer, and at the same time ensures long-term operation of the system. Frequent starting and stopping can place the system in a severe electrochemical environment. In particular, when the anode side of the fuel cell is filled with air, as would be the case in standby situation, and the system is started up with the flow of a hydrogen rich reactants through the anode, a hydrogen/air boundary forms on the anode side of the cell [61]. This boundary moves from the anode inlet to outlet, and in doing so sets up a high voltage ( $\approx 1.4$  V) situation across the cell. This potential sets up a reverse current which greatly accelerates the oxidation of carbon, which theoretically can proceed as:



In fact, this potential driven “corrosion” of carbon can be quite severe, causing substantial loss of the

electrochemically active surface area as the electrode degrades with the loss of the catalyst support. Enhanced fuel cell degradation can occur under the additional stress conditions associated with cold start and hot stopping [62].

In order to deal with the problem of carbon corrosion due to start-stops, it is necessary to understand the critical parameters associated with the start-stop process, as well as the critical specifications that need to be placed upon the materials. What is required to do this is to have a validated model of carbon corrosion and its dependence on the critical parameters: temperature, pressure, flow rate, and composition. Such a model has been developed, to the first order, by Meyers and Darling [63] in which the potentials driving the reaction are linked to the spatial concentrations of oxygen and hydrogen in the flow channels.

Carbon corrosion can also occur due to fuel starvation on the anode, which sometimes happens when flooding is imminent. Again, this situation sets up localized high potentials and reverse currents. Carbon as a catalyst support is problematic, and for this reason control systems need to provide algorithms to deal with the situation when a particular cell in the stack is behaving in an abnormal fashion (so-called stack health algorithms).

### Load Following

In most applications involving fuel cells, the load is dynamic, and it is a requirement that the fuel cell system be able to respond to these load transients nearly instantaneously. This requires careful system design. This requirement translates into the need for having rapid delivery of fuel to the catalyst sites within each cell in the stack. Two transient type of operating modes were described above, startup from subfreezing and start/stop. Here focus will be on load changes during operation, and on applications involving both hydrogen and reformate fuel. In the former, emphasis will be placed on the effects of rapid cycling on durability, whereas in the latter, emphasis will be placed upon achievement of the required critical parameters over the expected range in operating parameters.

Hydrogen-air fuel cells, which are used in transportation, backup power, and material handling applications, respond almost instantaneously, provided that they a fed adequate amounts fuel. Control schemes for

hydrogen-based fuel cell systems for automotive applications are discussed by Pukrushpan et al. [64] in which the primary focus of the discussion is on control of the supply of air and hydrogen. In some cases, the fuel cell system will be “hybrid” with a battery in the overall system architecture. For example, in telecom backup fuel cell systems, a small battery pack (or supercapacitor) is included to ensure the continuous delivery of power within the first millisecond after the grid drops out. In light duty vehicles, batteries are included as a means of load leveling and energy harvesting during breaking. In forklift applications, batteries offer the opportunity to alleviate the load-following requirements on the stack, thereby extending its operating life. In this regard, it is important from a systems controls perspective that one of the key aspects of achieving adequate load-following capability is to develop a comprehensive stack control scheme that incorporates the relevant system-level components, critical parameters, and their linkage to the critical specifications. In most cases, this control scheme must address both stack health and efficiency.

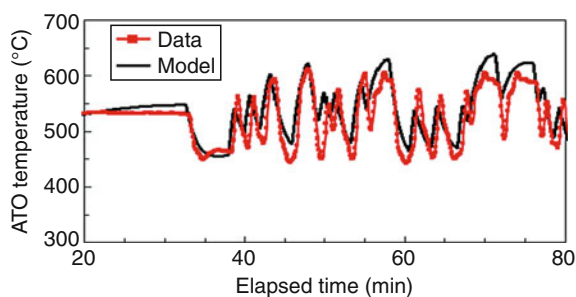
For reformate-based systems, the control problem becomes particularly difficult because of the strong interactions between subsystems. This is perhaps best seen when considering the function of the ATO. The ATO is designed to oxidize the unreacted anode gas into water for the reformer, as well as to achieve zero emissions from the fuel cell system. On the other hand, the ATO also serves as the heat source for steam generation. Both the stack and the ATO require an air/oxygen source. To minimize cost, improve efficiency, and increase the operating latitude over which the system can be water independent, it is desirable to use a single air blower to provide air to both the stack and the ATO. However, this shared actuator couples the controls for stack, ATO, steam generation and reforming, thereby increasing the complexity of the interaction among the control loops.

Another complexity arises due to the lack of in situ measurement of some of the critical parameters, requiring the inference of the value of the critical parameter from the operation of the system itself. This is particularly true when, by design, the relative humidity of the input reactant streams is near or at saturation. In this situation, RH sensors become unreliable and the RH level needs to be inferred from the system control

parameters. In order to accommodate the strong interactions between the subsystems, a considerable amount of modeling and simulating of system performance is required. This can be accomplished by linking process models in ASPEN Dynamics<sup>®</sup> with controls models in Matlab<sup>®</sup>. In this way, nonlinear dependencies in the process can be linearized in a piecewise fashion, and exported to MatLab, an environment in which the order of the model can be systematically reduced together with weighting factors to get time-dependent models that can provide unified descriptions of the fuel cell in both steady state and dynamic environments. An example of the result of such an exercise is shown below in Fig. 8, comparing the performance of the ATO during load transients versus the model predictions. In this particular experiment, the system was run in steady state, and the load-following requirements were simulated by suddenly manipulating the cathode air blower in pseudorandom manner. It is seen that the dynamic model is a remarkably good predictor of the ATO performance [65].

### CO Poisoning

One of the challenges facing the fuel cell system is the mismatch in the time constants of the various components in the system. The reformer has the longest time constant, followed by that of the air handling system and stack, and finally the battery. This mismatch can result in large fluctuations in the quality of fuel provided to the stack, and can be the source of unwanted excursions in the amount of CO reaching the stack. With validated advanced simulation tools, it is possible to develop robust control strategies that ensure that the



**Polymer Electrolyte (PE) Fuel Cell Systems. Figure 8** Model versus data comparison for transient operation of ATO [61]

overall system can meet the load-following requirements during large and rapid transients. Shown below in Fig. 9 is the transient response of the Plug Power ATR-based fuel processor to large and sudden step changes in electrical load demand. Here, it is seen the control software is able to keep the CO concentration within performance specifications (<10 ppm). In this case, batteries are used to make up the load demand while the reformer ramps up, illustrating the mismatch in time constants between the stack and reformer subsystems [66].

It is evident that advanced controls, once validated with system verification tests, can be used to maximize the system operational latitude. As shown below in Fig. 10, these controls are used to optimize the operating window for the CO concentration in the reformat stream fed to the stack. Here, it is seen that there is a large operational window, in terms of PROX outlet temperature and CO concentration over which the system can operate within its specification.

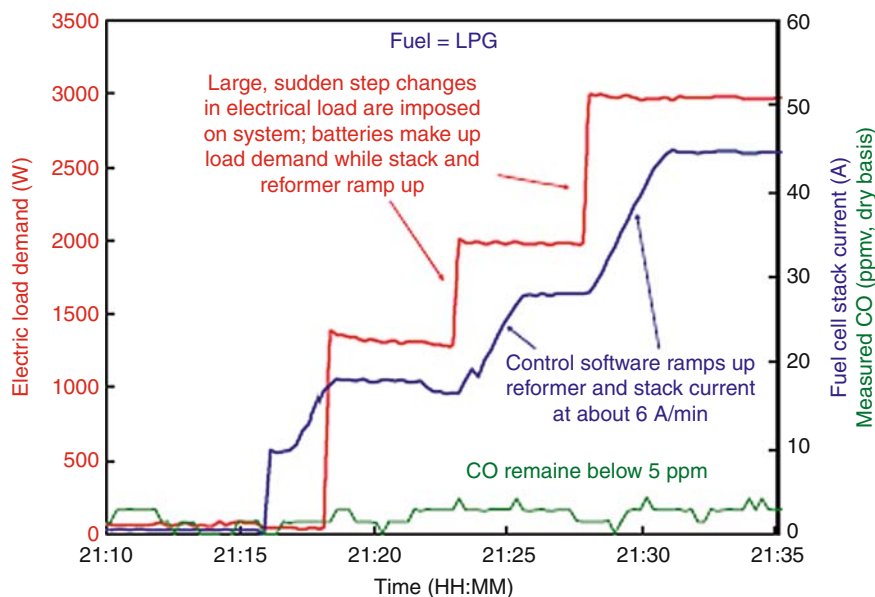
It is perhaps important at this point to once again discuss the potential advantages of higher temperature operation. It is well known that fuel cell systems that operate at higher temperatures can tolerate much higher levels of CO. Indeed, tests have shown that stacks at temperatures in the range of 160–200°C can tolerate CO concentrations on the order of  $10^3$ – $10^4$  ppm. In this case, it is possible to eliminate the PROX and its associated costs and sources of unreliability, while simplifying the controls requirement.

There are of course, many other potential failure modes that need to be addressed. Only a few are mentioned here. Perhaps, the most difficult are those associated with maintenance of the polymer electrolyte's required properties over hours of operation and over a wide range of conditions. Temperature, humidity, and potential cycling place severe stress on the polymer electrolyte and its matching electrodes, and minimizing these through great attention to engineering detail is fundamental to achievement of technical readiness.

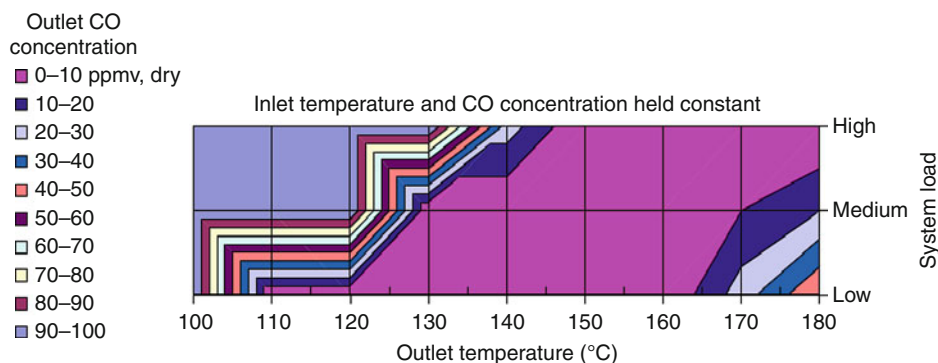
### System Cost and Reliability

Two key parameters that are obstacles to widespread fuel cell adoption are system cost and reliability. Both are interrelated through the system design, driven mainly by the complexity of the interactions between the stack and the rest of the system, which in turn is determined





**Polymer Electrolyte (PE) Fuel Cell Systems. Figure 9**  
 Fuel processor performance during load transients [61]



**Polymer Electrolyte (PE) Fuel Cell Systems. Figure 10**  
 Prox CO operating latitude in terms of temperature and load [61]

by the nature of the polymer electrolyte. However, system complexity is also determined by the customer-driven system specifications. System specifications can vary widely, depending upon the application, and for this reason it is nearly impossible to make specific claims about costs of fuel cell systems without reference to the detailed requirements driving the design.

The cost and reliability of PEM fuel cell systems for light vehicles are tracked quite closely as automotive manufacturers are making significant investments in developing fuel cell cars for field entry in the next

5–10 years. The US Department of Energy (DOE) has, in conjunction with industry leaders, developed cost and reliability performance targets and tracks progress against those targets in its annual review process. In addition, the DOE has defined a baseline fuel cell system with which to model performance improvements due to advances in technology and to coordinate those findings with the assumptions in the cost projections. In its most recent (2010) Annual Merit Review (AMR), stack cost projections indicate that at low production volumes the membrane dominates the

stack cost, whereas at high-production volumes, the catalyst ink dominates the cost. Cost estimates suggest that in high volume production, stack costs (based upon 3M NSTF nanostructured electrodes) of around \$25/kW<sub>e(net)</sub> can be achieved, and system costs of around \$51/kW<sub>e(net)</sub> can be achieved, nearly equal to the 2010 DOE targets [67].

The situation for stationary fuel cells is not as clear. The current 2011 DOE performance targets for integrated stationary fuel systems of \$750/kW<sub>e</sub>, 40,000 h of durability and 40% electrical efficiency at rated power are difficult targets to meet with a reformat-based system. These targets are under review and will most likely be updated and more clearly aligned with the intended application.

The system requirements for stationary systems can vary widely, depending upon application. In general, however, the key to cost reduction is through system simplification. One way to accomplish this is through good engineering design, including value analysis/value engineering (VA/VE). VA/VE enables one to reduce the number of parts. Another way is to avoid expensive secondary fastening operations by having parts and modules self-align to each other, eliminating adjustments, using gravity wisely, and designing for “Poka Yoke.”

Another way to reduce cost and improve reliability is through advances in technology. As mentioned above, the pursuit of polymer electrolytes that are more tolerant to low RH conditions, that do not require external humidification, and that run at higher temperatures is a prime example of how technology can reduce cost and complexity. Elimination of sensors through intelligent and advanced controls is another example.

Currently, stationary fuel cells include the cost burden of the reformer, which makes the achievement of \$750/kW<sub>e</sub> difficult. As a percentage of the total system cost, the reactant processing module can be as much as 30% of the total system unit manufacturing and service cost. Whereas the stack can scale downward gracefully as a function of size (kW<sub>e</sub> output), the reformer cannot, making low-cost small residential stationary systems even more challenging. For this reason, it is perhaps appropriate to examine *alternative system architectures* that decouple the reformer, and provide the required hydrogen from a shared-use centralized reformer. Such architectures are currently under consideration for use in Japan.

Lastly, it seems appropriate that at this stage fuel cell systems should be developed for large niche markets that are appropriate for the current state technical readiness of the technology. Forklift applications are a prime example of matching requirements versus capability. Large stationary fuel cells that are designed for constant base loads with high combined heat and power utilization requirements is another.

## Fuel Cell Systems and Sustainability

The connection between fuel cell systems and sustainability is of course, hydrogen. The lightest and most abundant element on planet Earth is destined to play a central role in the energy landscape of the future. Renewable energy sources can be used to produce hydrogen, which can then be stored as a means of addressing the natural intermittency of the renewable source. In the future, by splitting water with a renewable source of energy such as wind and solar energy, and using the stored hydrogen so produced to produce electricity, one can also return the product water to its source, achieving a sustainable energy paradigm for nearly every nation on earth. This is truly the energy “end game.” Indeed, the fact that as an energy carrier hydrogen can play a key role in the storage of energy produced from renewable resources has received renewed interest among a variety of stakeholders.

Since the fuel cell system will play a key role in the pursuit of sustainable energy, whether in electrifying the automobile for transportation, providing energy for buildings, or converting stored hydrogen generated from renewable sources into electricity, the design of the fuel cell itself should consider its impact on the environment. This encyclopedia addresses the subject of life cycle analysis elsewhere. Here, it seems appropriate to note that Cooper et al. [68] have developed a method to assist in the rapid preparation of life cycle assessments of energy generation technologies, and in particular polymer electrolyte fuel cell systems. The method allows one to compare different fuel cell system design approaches, using publically available and peer-reviewed Life Cycle Assessment data, against an environmental impact weighting scheme that reflects various environmental sensitivities. Providing a fuel cell system “designed for the environment,” with the

goal that no part ever ends up in a landfill, seems to be a fitting way to approach the objective of truly sustainable energy systems.

### Future Directions

Progress in the development of fuel cell systems for transportation, stationary, and motive power applications will continue as long as the incentives around renewable and clean energy continue to receive public support. This support, in turn, needs to be based upon public awareness of the true cost of the incumbent technologies: environmental and social, as well as financial. The rate of progress made in the past 10 years has been accelerating due to the support of governments tending to the longer range needs of their citizens. With a view of the need for a clean energy future clearly in sight, alternative means of energy generation, conversion, and storage will play an increasing role in shaping the future for our children's children, a time not too far away.

To meet this challenge, fuel cell research and development will need to continue to pursue a range of subjects from basic understanding of the key factors governing charge transfer at the nanoscale to enabling the simplification of systems by developing advanced materials that can relax the constraints currently imposed by the polymer electrolyte membrane. Hence, continued examination of the property and structure relationships in electrocatalysis will be needed to point the way to improved catalyst materials. Research on improved and more stable nanostructured electrodes incorporating advanced and lower cost materials will be needed to lower cost while improving durability. The development of new polymer electrolyte membrane materials and structures with broader operating latitudes at higher operating temperatures will be needed to further reduce cost by enabling less complex system designs. In parallel with this, of course, innovative engineering approaches toward reducing cost and increasing durability will be needed to close the gap between today's cost-performance curves and those required to enable penetration of commercial markets. In other words, a broadly based and a steady attack on the issues of cost and durability is necessary to advance the state of the art around PEM systems.

### Bibliography

1. Barbir F (2005) PEM fuel cells: theory and practice. Elsevier, Burlington
2. Srinivasan S (2006) Fuel cells: from fundamentals to applications. Springer, New York
3. Vielstich W, Lamm A, Gasteiger HA (eds) (2003) Handbook of fuel cells. Wiley, Chichester
4. Garland N, Kopasz JP (2007) The United States Department of Energy's high temperature, low humidity membrane program. *J Power Sources* 172:94–99
5. Mader J, Xioa L, Schmidt TJ, Benicewicz BC (2008) Polybenzimidazole/acid complexes as high temperature membrane. In: Fuel cells: advances in polymer science, vol 216. Springer, Berlin, pp 63–124
6. Vishnyakov A, Niemark AV (2000) Molecular study of Nafion membrane solvation in water and methanol. *J Phys Chem B* 104:4471–4478
7. Kreuer DK, Paddison S, Spohr E, Schuster M (2004) Transport in proton conductors for fuel cell applications: simulations, elementary reactions, and phenomenology. *Chem Rev* 104:4637–4678
8. Springer TE, Zawodzinski TA, Gottesfeld S (1991) Polymer electrolyte fuel cell model. *J Electrochem Soc* 138:2334–2341
9. Fimrite J, Carnes B, Struchtrup H, Djilali N (2009) Coupled proton and water management in polymer electrolyte membranes. In: Paddison SJ, Promislow KS (eds) Device and materials modeling in PEM fuel cells, vol 113, Series: topics in applied physics. Springer, New York
10. Weber AZ, Newman J (2004) Modeling transport in polymer-electrolyte fuel cells. *Chem Rev* 104:4679–4726
11. Wang CY (2004) Fundamental models for fuel cell engineering. *Chem Rev* 104:4727–4765
12. Darling RM, Meyers JP (2003) Kinetic model of platinum dissolution in PEMFCs. *J Electrochem Soc* 150:A1523–A1527
13. Meyers JP, Darling RM (2006) Model of carbon corrosion in PEM fuel cells. *J Electrochem Soc* 155:A1432–A1442
14. Reiser CA, Bregoli L, Patterson TW, Yi JS, Yanag JD, Perry MI, Jarvi TD (2005) A reverse current decay mechanism for fuel cells. *Electrochem Solid State Lett* 8(6):A273–A276
15. [www.cd-adapco.com](http://www.cd-adapco.com)
16. [www.fluent.com](http://www.fluent.com)
17. Markovic N, Gasteiger H, Ross PN (1997) Kinetics of oxygen reduction on Pt (hkl) electrodes: implications of the crystalline size effect with supported Pt electrocatalysts. *J Electrochem Soc* 144:1591–1597
18. Nørskov JK, Rossmeisl J, Logadottir A, Lindqvist L, Kitchin JR, Bligaard T, Jónsson H (2004) Origin of the overpotential for oxygen reduction at a fuel cell cathode. *J Phys Chem B* 108:17886–17892
19. Mavrikakis M, Hammer B, Nørskov JK (1998) Effect of strain on the reactivity of metal surfaces. *Phys Rev Lett* 81:2819–2822
20. Taylor CD, Wasileski SA, Filhol JS, Neurock M (2006) First principles modeling of the electrochemical interface: consideration and calculation of a tunable surface potential from atomic and electronic structure. *Phys Rev B* 73:65402

21. Adachi H, Ahmed S, Lee SHD, Papadias D, Ahluwaia RK, Bendert JC, Adzic RR et al (2007) Platinum monolayer fuel cell electrocatalysis. *Top Catal* 46:249
22. Stamenkovic V, Mun BS, Mayrhofer KJJ, Ross PN, Markovic NM, Rossmeisl J, Greeley J, Nørskov JK (2006) Changing the activity of electrocatalyst for oxygen reduction by tuning the surface electronic structure. *Angew Chem Int Ed* 45:2897
23. Nørskov JK, Rossmeisl J, Logadottir A, Lindqvist L, Kitchin JR, Bligaard T, Jonsson H (2004) Origin of the overpotential for oxygen reduction at a fuel cell cathode. *J Phys Chem B* 108:17886–17892
24. Koper MTM (ed) (2009) Fuel cell catalysis. Wiley, Hoboken
25. Gasteiger HA, Kocha SS, Sompali B, Wagner FT (2005) Activity Benchmarks for Pt, Pt-alloy and non-Pt oxygen reduction catalysts for PEMFCs. *Appl Catal B* 56:9–35
26. Debe MK (2003) Novel catalysts, catalyst support and catalyst coated membrane methods. In: Vielstich MW, Lamm A, Gasteiger HA (eds) *Handbook of fuel cells-fundamentals, technology and applications*, vol 3. Wiley, New York, Chapter 45
27. Gancs L, Kobayashi T, Debe MK, Atanasoski R, Wieckowski A (2008) Crystallographic characteristics of nanostructured thin-film fuel cell electrocatalysts: a HRTEM study. *Chem Mater* 20(7):2444–2454
28. Debe MK (2010) Advanced cathode catalysts and supports for PEM fuel cells. 2010 DOE Hydrogen Program Merit Review and Peer Evaluation, Washington, DC, 11 June 2010
29. Tang JM, Jensen K, Waje M, Li W, Larsen P, Pauley K, Chen Z, Ramesh P, Itkis M, Yan Y, Haddon R (2007) High performance hydrogen fuel cells with ultralow Pt loading carbon nanotube thin film catalysts. *J Phys Chem C* 111(48):17901–17904
30. Antonlini E, Gonzalez ER (2009) Ceramic Materials as supports for low temperature fuel cell catalysts. *Solid State Ionics* 180:746–763
31. Scherer GG (ed) (2008) Fuel cells II, vol 216, *Advances in polymer science*. Springer, Berlin
32. Abhishek R, Yu X, Dunn S, McGrath JE (2009) Influence of microstructure and chemical composition on proton exchange membrane properties of sulfonated-fluorinated hydrophilic-hydrophobic multiblock copolymers. *J Membr Sci* 327(1–2):118–124
33. Lipp L (2010) High temperature membrane with humidification independent cluster structure. 2010 DOE Hydrogen Program Merit Review and Peer Evaluation, Washington, DC, 11 June 2010
34. Inaba M (2009) Chemical degradation of perfluorinated sulfonic acid membranes. In: Buechi FN, Inaba M, Schmidt TJ (eds) *Polymer electrolyte fuel cell durability*. Springer, New York
35. Danilczuk M, Coms FD, Schlick S (2008) Fragmentation of fluorinated model compounds exposed to oxygen radicals: spin trapping ESR experiments and implications for the behaviour of proton exchange membranes used in fuel cells. *Fuel Cells* 8(6):436–452
36. Ghassemzadeh L, Marrony M, Barrera R, Kreuer KD, Maier J, Mueller K (2009) *J Power Sources* 186(2):334–338
37. Antoine O, Durand R (2000) *J Appl Electrochem* 30:839–844
38. Liu H, Gasteiger HA, Laconti A, Zhang J (2006) *Electrochemical Soc Trans* 1(8):283–293
39. Mittal V, Kunz R, Fenton JM (2006) *Electrochemical Soc Trans* 1(8):275–282
40. Miyake N, Wakizoe M, Honda E, Ohta T (2006) *Electrochemical Soc Trans* 1(8):249–261
41. Yu J, Matsuura T, Yoshikawa Y, Islam MN, Hori M (2005) *Electrochem Solid State Lett* 8:A156–A158
42. Curtin DE, Lousenberg RD, Henry TJ, Tangeman PC, Tisack ME (2004) *J Power Sources* 131:41–48
43. Lassegues JC, Schoolmann D, Trinquet O (1992) Proton conducting acid polymer blends. In: Balkanski T, Takahashi T, Tuller HL (eds) *Solid State Ionics*. Elsevier, Amsterdam, pp 443–448
44. Lassegues JC (1992) Mixed inorganic-organic systems: the acid/polymer blends. In: Colomban PH (ed) *Proton conductors: solids, membranes and gel – materials and devices*. Cambridge University Press, Cambridge, pp 311–328
45. Savinell R, Yeager E, Tryk D, Landau U, Wainright J, Weng D, Lux K, Litt M, Rogers C (1994) A polymer electrolyte for operation at temperatures up to 200°C. *J Electrochem Soc* 141(4):L46–L48
46. Schmidt TJ (2009) High-temperature polymer electrolyte fuel cells: durability insights. In: Büchi FN, Inaba M, Schmidt TJ (eds) *Polymer electrolyte fuel cell durability*. Springer, New York
47. Neyerlin KC, Singh A, Chu D (2008) Kinetic Characterization of a Pt-Ni/C catalyst with a phosphoric acid doped PBI membrane in a proton exchange membrane fuel cell. *J Power Sources* 176:112–117
48. Gottesfeld S, private communication (2010)
49. McGrath MF, Anthony MT, Shapiro AR (1992) *Product development*. Butterworth-Heinemann, Boston
50. Patterson ML, Fenoglio JA (1999) *Leading product innovation*. Wiley, New York
51. McGrath M (2001) *Product strategy*. McGraw-Hill, New York
52. Du B, Pollard R, Elter JF, Ramani M (2009) Performance and durability of a polymer electrolyte fuel cell operating with reformat: effects of CO, CO<sub>2</sub>, and other trace impurities. In: Büchi F, Inaba M, Schmidt TJ (eds) *Polymer electrolyte fuel cell durability*. Springer, New York, pp 341–366
53. Du B, Pollard R, Elter JF (2006) CO-air bleed interaction and performance degradation study in proton exchange membrane fuel cells. *Electrochem Soc Trans* 3(1):705–713
54. Adachi H, Ahmed S, Lee SHD, Papadias D, Ahluwaia RK, Bendert JC, Kanner SA, Yamazaki Y (2009) A natural gas fuel processor for a residential fuel cell system. *J Power Sources* 168:244–255
55. Feitelberg AS, Rohr DF Jr (2005) Operating line analysis of fuel processors for PEM fuel cell systems. *Int J Hydrogen Energy* 30:1251–1257
56. Feitelberg A (2003) On the efficiency of PEM fuel cell systems and fuel processors. Presentation at fuel cell seminar, Miami, 3 November 2003
57. Mukundan R, Kim YS, Garzon F, Pivovar B (2006) Freeze/thaw effects in PEM fuel cells. *ECS Trans* 1:403–413

58. Guo QH, Qi ZH (2006) Effect of freeze-thaw cycles on the properties and performance of membrane-electrode assemblies. *J Power Sources* 160:1269–1274
59. Lee C, Merida W (2007) Gas diffusion layer durability under steady state and freezing conditions. *J Power Sources* 164:141–153
60. Mao L, Wang C-Y (2007) *J Electrochem Soc* 154:B139–B146
61. Tang H, Qi Z, Ramani M, Elter JF (2006) PEM Fuel cell cathode carbon corrosion due to the formation of air/fuel boundary at the anode. *J Power Sources* 158:1306–1312
62. Du B, Pollard R, Ramani M, Graney P, Elter JF (2007) Impact of cold start and hot stop on the performance and durability of a proton exchange membrane (PEM) fuel cell. *ECS Trans* 5(1):271–282
63. Meyers JP, Darling RM (2006) Model of carbon corrosion in PEM fuel cells. *J Electrochem Soc* 153(8):A1432–A1442
64. Pukrushpan JT, Stefanopolou AG, Peng H (2005) Control of fuel cell power systems. Springer, London
65. Feitelberg AS, Elter JF (2005) Development, design and performance of plug power's next generation stationary PEM fuel cell system prototype. Presented at 2005 fuel cell seminar, Palm Springs, California, 14–18 November 2005
66. Elter JF (2007) The design and control of fuel cell systems. Presented at the H<sub>2</sub> fuel cells millennium convergence, Bucharest, Romania, 21, 22 September 2007
67. James B (2010) Mass-production cost estimation for automotive fuel cell systems. DOE Annual Merit Review, Washington, DC, June 2010. [www.annualmeritreview.energy.gov/](http://www.annualmeritreview.energy.gov/)
68. Cooper J, Lee S-J, Elter J, Boussu J, Boman S (2009) Life cycle design metrics for energy generation technologies: method, data and case study. *J Power Sources* 186:138–157

## Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications

SHYAM S. KOCHA

Hydrogen Technologies and Systems Center, National Renewable Energy Laboratory, Golden, CO, USA  
 Fuel Cell Laboratory, Nissan Technical Center North America, Farmington Hills, MI, USA

### Article Outline

Glossary

Definition of Subject: Automotive PEM Fuel Cells

Introduction

Automotive PEMFCs

Future Directions

Bibliography

### Glossary

**Automotive PEMFC** Proton exchange membrane fuel cell stacks used to power automotive vehicles typically using hydrogen as a fuel and ambient air as the oxidant.

**Electrocatalyst** The material used on the anode and cathode electrodes of fuel cells to catalyze the fuel oxidation and oxygen reduction reactions to produce electrical power and by-products of heat and water. Amount of electrocatalyst used in the anode or cathode of fuel cells is reported in units of mg/cm<sup>2</sup>.

**Membrane/PEM** The proton conductive polymer electrolyte used to separate the anode and cathode compartments of fuel cells. The membrane replaces the liquid electrolytes used in some fuel cells.

**Fuel cell performance** The voltage produced by a fuel cell stack at a defined current density. A performance or polarization curve refers to a plot of the cell potential (*V*) versus current density (*I*) under specified conditions of pressure, temperature, humidity, and reactant stoichiometry.

**Fuel cell durability** A measure of the degradation of components of a fuel cell as well as the output power of the entire stack over time. Also defined in terms of the maximum life of the stack before failure or degradation rate of the fuel cell performance in μV/h.

### Definition of Subject: Automotive PEM Fuel Cells

Since the discovery of fuel cells in the nineteenth century, they have been designed for operation with liquid alkaline, acid, and solid oxide ion conducting electrolytes in different temperature ranges to produce electrical power for stationary, portable, and automotive applications. The liquid acid that provides ionic conduction has been replaced by fairly thin proton conducting membranes such as polystyrenes and perfluorosulfonic acids (PFSA) like Nafion and more recently with hydrocarbon-based polymers. These fuel cells incorporating a proton-conducting membrane rather than liquid electrolyte to separate the anode and cathode (forming a 3-layer sandwich or catalyst coated membrane) are referred to as PEMFCs. PEMFCs are preferred for use in automobiles for a multitude of reasons including their high volumetric and gravimetric power density.

PEMFCs for automotives have electrodes that are typically constituted of Pt-based catalysts separated by proton-conducting perfluorosulfonic acid (PFSA) or hydrocarbon membranes. The membranes ( $\sim 25 \mu\text{m}$  thick) have proton conductivities of about  $100 \text{ mS/cm}$  and areal resistances of  $50 \text{ m}\Omega\text{-cm}^2$ . The anode Pt loadings that catalyze the hydrogen oxidation reaction (HOR) are of the order of  $0.05 \text{ mg/cm}^2$  while the cathode Pt loadings that catalyze the oxygen reduction reaction (ORR) fall in the range  $0.20\text{--}0.40 \text{ mg/cm}^2$ . Automotive PEMFCs are operated in the temperature range from ambient to  $\sim 90^\circ\text{C}$  at ambient to  $\sim 300 \text{ kPa}$  and at  $30\text{--}100\%$  RH. Compressed hydrogen fuel tanks ( $350\text{--}700 \text{ kPa}$ ) and ambient air pressurized using a compressor are employed as the fuel and oxidant sources. Depending on the size of the vehicle, automotive fuel cell stacks produce  $80\text{--}140 \text{ kW}$  of peak power. Most fuel cell automotives are hybrids and employ a  $10\text{--}20 \text{ kW}$  NiMH or Li ion battery to improve efficiency and to store and provide supplemental power. Automotive PEMFCs are subject to variable operating conditions of high potentials, load cycling, start-up and shutdown cycles, humidity cycles, freeze-thaw cycles, and contamination from ambient air. The main obstacles toward commercialization of PEMFC stacks for automotives are the combination of cost, performance, and durability that are not mutually exclusive. Today's PEMFC-powered automobiles demonstrate driving ranges and lifetimes approaching ( $\sim 70\%$ ) that of ICE vehicles. Major automotive companies have stated that PEMFCs for automotives are slated to arrive at cost levels approaching that necessary for commercialization beginning in 2015.

Currently, ICE-powered vehicles emit  $\sim 1.5$  billion tons  $\text{CO}_2$  equivalent per year at a US urban air pollution cost to society of  $\$30$  billion/year. The consumption of fossil fuels by the human species ( $\sim 6.5$  billion) has resulted in challenges of energy sustainability, environmental pollution, and global warming that need to be addressed urgently. Currently, several technologies that lower the green house gas emissions partially such as gasoline-powered hybrid electric vehicles (HEVs) and gasoline plug-in hybrids (PHEVs), biofuel PHEVs, and batteries (BEVs) are being developed in parallel with fuel cells.

PEMFCs can operate on hydrogen fuel and atmospheric air to produce electrical energy, while

exhausting only heat and water. Hydrogen is not available on earth as gas; it is found as a compound bound to oxygen as in water or bound to carbon, and in living things as biomass. Hydrogen is a carrier of energy and needs to be generated and stored efficiently; an infrastructure for hydrogen needs to be developed along with more efficient storage of hydrogen carried on board the vehicle. Currently about 9 million metric tons of hydrogen per year are generated in the USA that could power 30 million automobiles. Hydrogen can be produced from fossil fuels such as natural gas, and also, renewable sources, such as hydroelectric, wind, geothermal, solar photovoltaics, direct photoelectrochemical, concentrated solar power ocean (tidal, wave, current, and thermal), etc. The application of  $\text{H}_2/\text{Air}$  PEMFCs in automotives is one of the most important components in a renewable hydrogen economy that has the potential to reduce greenhouse gas emissions (to  $80\%$  below 1990 GHG levels), lower pollution, and arrest global warming.

## Introduction

Fuel cells have been known to science for more than 150 years. In 1800, British scientists William Nicholson and Anthony Carlisle first demonstrated and explained the phenomena of using electricity to decompose water into hydrogen and oxygen. William Robert Grove (1811–1896), a Welsh scientist who was working on electrolysis of water to hydrogen and oxygen, tested the hypothesis that the reverse might be possible. He placed two platinum strips immersed in dilute sulfuric acid in two separate chambers, one of which was filled with hydrogen and the other with oxygen. A current was found to flow between the two platinum strips and water produced in the chamber confirming the hypothesis [1]. Although Grove was the first to build a working fuel cell, the discovery of the principle and fundamentals of the fuel cell is attributed to Christian Friedrich Schoenbein (1799–1868) [2]. Grove later improved on his original experiments by using a series of four cells to increase the total voltage; he named the device a “gas battery” – now known as a fuel cell stack.

Significant contributions were made in later years on fuel cells powered by various fuels such as that by Mond and Langer [3], Haber, W. V. Jacques [4], Bauer [5], Taitelbaum, Schmid [6], Tobler [7] and others.

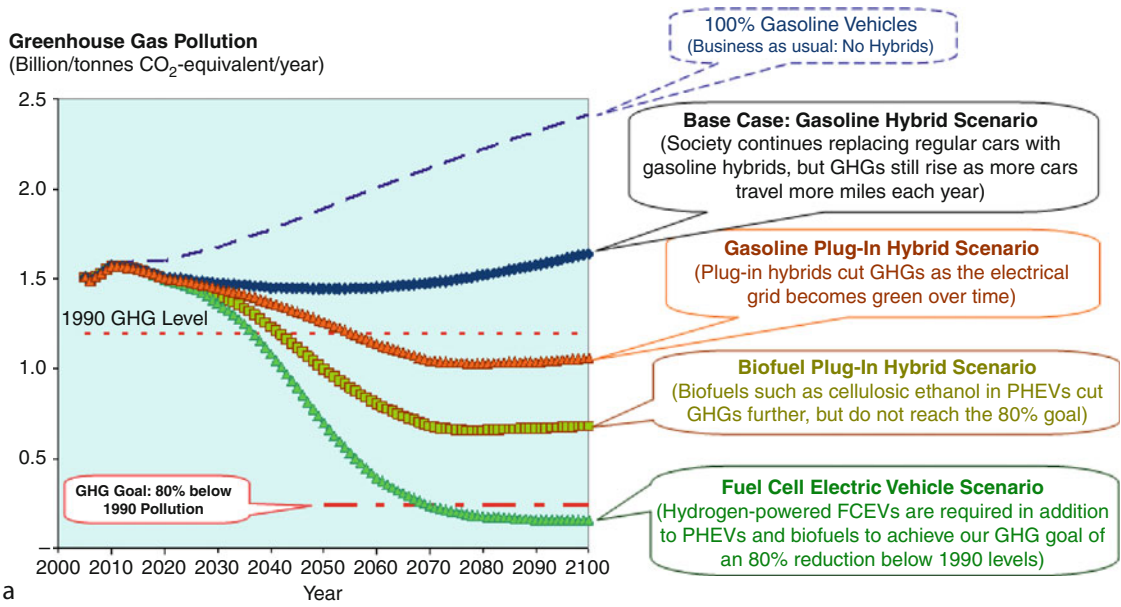
In addition to the experimental and practical fuel cell devices, a number of scientists contributed to the science and mechanism of the detailed functioning of fuel cells. Two noteworthy theories were debated to explain the functioning of the fuel cell. One was the “contact” theory originally proposed by Alessandro Volta (1745–1827) to explain his battery and the other was a “chemical” theory that held chemical reactions responsible for the generated power. Both theories had part of the solution which is that reactions that occurred where reactant gases, Pt catalyst, and electrolyte converged. This understanding was advanced by the contributions of Friedrich Wilhelm Ostwald (1853–1932). Other systematic nonempirical contributions were made by Nernst [8], Tafel [9], Erdey-Gruz and Volmer. A history of the development of fuel cell electrodes between 1839 and 1960 can be found in the reviews of Liebhasvsky and Cairns [10], Vielstich [11], Baur [5], Tobler [7], Maget [12], and Liebhasvsky and Grubb [13].

In 1932, Francis Thomas Bacon replaced the platinum electrodes with cheaper porous nickel metal and the sulfuric acid with potassium hydroxide to demonstrate the first alkaline fuel cell (AFC). Alkaline fuel cells were demonstrated to produce power for practical applications such as welding machines, tractors, powerlifts, etc., in the 1950s. Improved AFCs (2.3 kW) were engineered by Pratt & Whitney/International Fuel Cells (IFC) and were used by NASA in manned US Apollo space missions (1968–1972) and Skylabs for about 54 missions. High power density AFCs using precious metal-based catalysts, static electrolyte (KOH), and operating on hydrogen and oxygen have been used to power (three 12 kW stacks, 92°C, 400 kPa) space shuttles since 1981 [14]. Other companies that have worked on AFCs include Union Carbide Corp., Siemens AG, and the European Space Agency [15].

Cation exchange resins polymerized as sulfonic acids became available in 1945 for use as de-ionizers. Around 1959, General Electric (Thomas Grubb and Leonard Niedrach) considered the use of these materials (sulfonated polystyrene) to form solid polymer electrodes (SPEs) as membranes for fuel cells; these materials were predicted to eliminate the system complexity involved in using liquid electrolytes and lead to the first PEMFCs [10, 16, 17]. The commercial availability of Teflon<sup>®</sup> (discovered in 1938 by Roy Plunkett

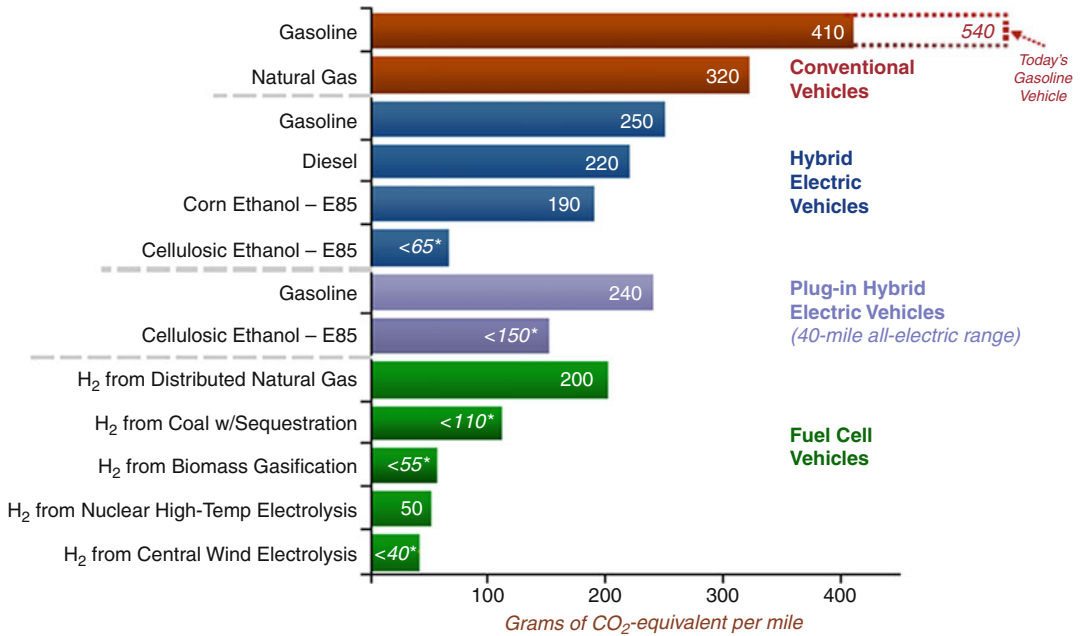
of Dupont<sup>®</sup>) enhanced the performance of SPEs due to its hydrophobic nature and consequent lowered flooding of electrode pores. Using SPEs, PEM modules were fabricated and used in the Gemini space modules operating under hydrogen and oxygen. The Biosatellite 2 (1967) followed the Gemini program (7 flights, 1962–1966) in which Nafion membrane was used for the first time. GE continued working on PEMFCs and in the 1970s developed PEM water electrolysis technology for undersea life support that was used in US Navy oxygen-generating plants. Some of the technology from GE was acquired by UTC affiliates Hamilton-Standard and IFC in 1984. The British Royal Navy also adopted this technology in the 1980s for their submarine fleet. Siemens A.G. commenced a fuel cell research program on AFCs and in 1984 implemented a 100 kW fuel cell in a German navy submarine; they also have PEMFCs installed in submarines operating with rated power of 34–120 kW with technology that allows a high power density and good thermal management. The efforts described above in all these related areas established the basis and provided the foundation and grounds for pursuing fuel cells for automotive applications.

In 2009, ICE vehicles consumed about 3.5 billion barrels/year gasoline and emitted 1.5 billion tons CO<sub>2</sub> equivalent per year of the greenhouse gases [18]. Figure 1a illustrates the CO<sub>2</sub> emissions over the next century for five different scenarios and Fig. 1b illustrates the well-wheels green house emissions projected for 2010 [19]. In the USA, 28% of the total energy used powers the transportation sector. The fuels commonly used for transportation are gasoline (62%), diesel (22%), jet fuel (9%), natural gas (2%). The by-products (that have an impact on the environment and human health) of petroleum products include CO<sub>2</sub>, CO, SO<sub>2</sub>, NO<sub>x</sub>, volatile organic compounds (VOCs), fine particulate matter (fine PM), lead, benzene, formaldehyde, acetaldehyde, 1,3 butadiene, etc. Automotives produce a large proportion of the pollution and greenhouse gases on earth; with global warming becoming an accepted reality, governments and automakers are finally making an effort to lower green house gas emissions. Although a number of intermediate low-emission technologies are being unravelled, PEMFCs are considered to be the best long-term solution since they can operate on hydrogen



a

**Well-to-Wheels Greenhouse Gas Emissions**  
(life cycle emissions, based on a projected state of the technologies in 2020)



\*Net emissions from these pathways will be lower if these figures are adjusted to include:

- \* The displacement of emissions from grid power—generation that will occur when surplus electricity is co-produced with cellulosic ethanol
- \* The displacement of emissions from grid power—generation that may occur if electricity is co-produced with hydrogen in the biomass and coal pathways, and if surplus wind power is generated in the wind-to-hydrogen pathway

b

\* Carbon dioxide sequestration in the biomass-to-hydrogen process

**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 1**

(a) CO<sub>2</sub> emissions over the next century for five scenarios where gasoline vehicles, gasoline-hybrid, gasoline plug-in hybrids, biofuel plug-in hybrids, and fuel cell electric vehicles are used as the power source [18]; (b) well-wheels greenhouse gas emissions based on state of technology in 2020 [19]



and atmospheric air to produce electrical energy, while exhausting heat and water with zero emissions.

Although, onboard reforming of methanol, gasoline, etc., was seriously considered and attempted, fitting a miniature chemical plant in the limited space of a fuel cell vehicle was found to be untenable. Hydrogen was unanimously selected as the choice of onboard fuel for automotive vehicles in the early twenty-first century with an understanding of its strengths and weaknesses. Hydrogen is not available on earth as gas; it is found as a compound bound to oxygen as in water or bound to carbon, and in living things as biomass. Hydrogen has the highest energy content of any fuel by weight (hydrogen: 143 MJ/kg, gasoline: 43 MJ/kg) and the lowest by volume. Hydrogen can be produced from fossil fuels such as natural gas, and renewable sources such as hydroelectric, wind, geothermal, solar photovoltaics, direct photo-electrochemical, concentrated solar power ocean (tidal, wave, current, and thermal), etc. Hydrogen is also classified based on criteria such as primary energy sources (hydro, nuclear, wind, solar, natural gas, etc.); methods of production (reforming, electrolysis, etc.); renewable/nonrenewable, etc. Currently about 9 million metric tons of hydrogen per year is generated (~95%—steam methane reforming (SMR) and the rest electrolysis); in the USA, this could in principle, power ~30 million automobiles.

Historically, there have been concerns about the safety of hydrogen, but in actuality, it is as safe and even safer than other flammable fuels such as gasoline and natural gas. A primary safety advantage of hydrogen is that it has very high diffusivity so that it dilutes rapidly to a nonflammable concentration in a reasonably ventilated space. Additionally, due to the absence of carbon and presence of water vapor in the combustion products of hydrogen, hydrogen fires release less radiant energy thus lowering incidents of secondary fires. The flammability limits, explosion limits, ignition energy, flame temperature, and stoichiometric mixture most easily ignited in air are all well documented for hydrogen. Codes and standards have been established for the safe use of hydrogen. Fuel cell vehicles are subjected to the same safety tests and crash/impact tests as gasoline-powered vehicles and have little trouble passing them.

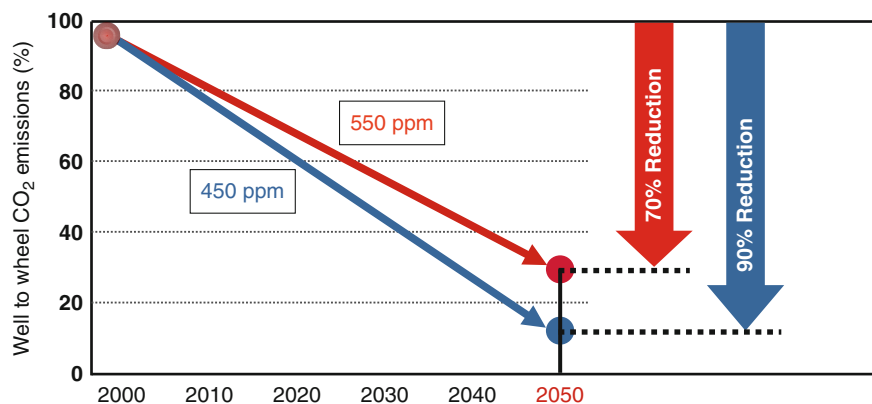
As the world transitions to a hydrogen economy, a hydrogen infrastructure including a combination of

distributed and centralized production is likely to evolve. Hydrogen pipeline networks already exist in some regions, often to provide hydrogen to the refining and food processing industry; transport by trucks is also prevalent. In the USA, at this time, there are 60 hydrogen fueling stations (~350 world-wide), 1,200 miles of hydrogen pipelines, and ~9 million tons of hydrogen produced every year. Hydrogen storage is often categorized as physical (or molecular) and chemical (or dissociative) storage. Onboard physical storage methods include compressed gas, liquid hydrogen, and cryo-adsorbed hydrogen; chemical storage includes metal hydrides and liquid organic carriers. Typically, compressed hydrogen (35–70 MPa) in one or two tanks (~4–8 kg hydrogen depending on the target range) is stored onboard fuel cell vehicles today. Most fuel cell vehicles today closely meet the driving range of their IC engine counterparts.

The application of H<sub>2</sub>|Air PEMFCs in automobiles is one of the most important components in a renewable hydrogen economy that has the enormous potential to reduce greenhouse gas emissions (to 80% below 1990 GHG levels) and arrest global warming. Based on the Intergovernmental Panel for Climate Change (IPCC) study (Fig. 2) by 2050 well-wheels emissions of CO<sub>2</sub> must be reduced by 70% ~90% versus 2000 levels [20]. Of all the possible pathways such as gas-electric hybrids, EVs, and fuel cells, only H<sub>2</sub>|Air PEMFCs have the ability to reduce emissions to zero. Thus in attempts to reduce greenhouse gas emissions, lower dependence on imported oil, and limit urban air and water pollution, FCVs are expected to play a very central role.

### Brief History of Automotive Fuel Cells

Ballard should be recognized for re-igniting the interest in PEMFCs in the late 1980s and 1990s and for the development of improved stacks used today by several companies in their fuel cell vehicles [21]. All the major automotive companies initiated fuel cell research and development programs as well as small-scale demonstration programs between 1998–2010; this was augmented by materials and component development by companies in the area of membranes, catalysts, diffusion media, bipolar plates, etc. In 2003, the US government announced a \$1.2 billion FreedomCAR



**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 2**

Reduction in well-to-wheel CO<sub>2</sub> emissions recommended for 2050 by the Intergovernmental Panel for Climate Change (IPCC) study [20]

and Hydrogen Fuel Initiative (HFI) to develop hydrogen-powered fuel cells. Together, by 2008, the two initiatives invested about \$1 billion to develop hydrogen-powered fuel cells, hydrogen infrastructure, and advanced automotive technologies. Fuel cells suffered a setback due to the 2009 US government policy that projected a longer 10–20-year forecast for automotive fuel cell commercialization. Nevertheless, most of the US, European, and Japanese automakers continue to support the development of PEMFCs for automobiles internally as well as in public statements and most of the government funding was later restored. Also, in 2009, the European Union (EU) announced €140 million (\$195 million) in available investments for research in energy technology. The funding (European Commission matched by contributions from the private sector) is part of a €1 billion (\$1.4 billion) that the EU plans to invest in fuel cell research and development by 2014. In Japan, the Ministry of Economy, Trade and Industry (METI)'s New Energy and Industrial Technology Development Organization (NEDO) has overseen a lot of the funding for fuel cell and hydrogen research, development, and demonstration. The Japan Hydrogen & Fuel Cell Demonstration Project (JHFC) conducts research and activities for the practical use of fuel cell vehicles and hydrogen stations. The JHFC consists of the Fuel Cell Vehicle-Demonstration Study and the Hydrogen Infrastructures-Demonstration Study; the studies are subsidized by the METI. Many countries now have hydrogen corridors or hydrogen highways with a

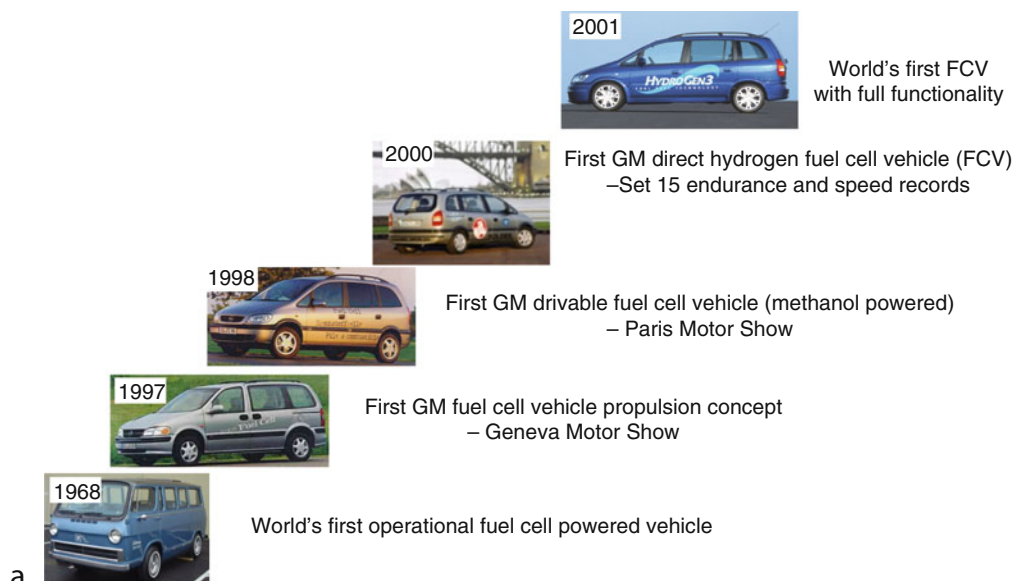
number of hydrogen fueling stations already implemented. In this section, for simplicity, the contributions made by a number of stack developers and automotive companies to advance fuel cell technology for automobiles based on publicly available sources will be outlined.

**Allis-Chalmers** In 1959, a team led by Harry Ihrig developed and demonstrated a 15 kW fuel cell tractor for Allis-Chalmers that was exhibited at state fairs across the USA. The FC stack system used KOH electrolyte and compressed hydrogen and oxygen as the reactants operating at 65°C. The original fuel cell tractor is on display at the Smithsonian. In 1965, Allis-Chalmers further developed hydrogen-powered FC golf carts. Allis-Chalmers is also experimenting with fuel cell stacks to generate power for spot welders and forklift trucks.

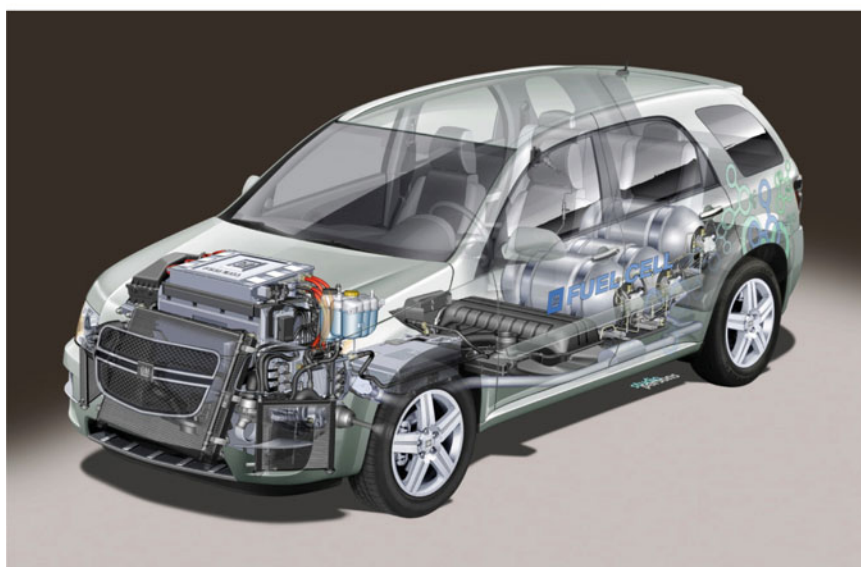
**General Motors** In 1968, GM's Electrovan was the automotive industry's first attempt at an automobile powered by a H<sub>2</sub>O<sub>2</sub> 125 kW fuel cell. It took a team of 250 people 2 years to demonstrate the potential feasibility of fuel cell technology. The Electrovan, weighed more than twice as much as a normal van and could travel at speeds up to ~70 mph for 30 s. In the early 1980s Los Alamos National Lab (LANL) initiated a PEMFC program and GM was part of the overview board. In 1996, the program moved from LANL to Rochester, NY, and later to Honeoye Fall, NY, where

a large fuel cell R&D center still operates. In 1997, GAPC was formed globally at Honeoye Falls, NY, Mainz Kastel, Germany, Warren, MI, and Torrance, CA. In 1998, the GM/DOE program was successful in demonstrating a 50 kW methanol fuel processor PEM system and in late 1998 – GM methanol Zafira was displayed at Geneva Auto Show. In 2000, focus shifted to hydrogen fuel and in 2001, HydroGen 3, Zafira was launched. Figure 3 illustrates the history of the various

GM FC vehicles from 1968 to 2001 and their latest FCV model. The advancements in volumetric and gravimetric power density of PEMFC stacks designed at General Motors between 1997 and 2004 are depicted in Fig. 4 [22]. Figure 5 is a photograph of their fifth generation stack [23]. Most recently, GM’s Chevrolet Equinox FCV passed 1 million miles with customers using the vehicles in everyday real-world conditions [24, 25].



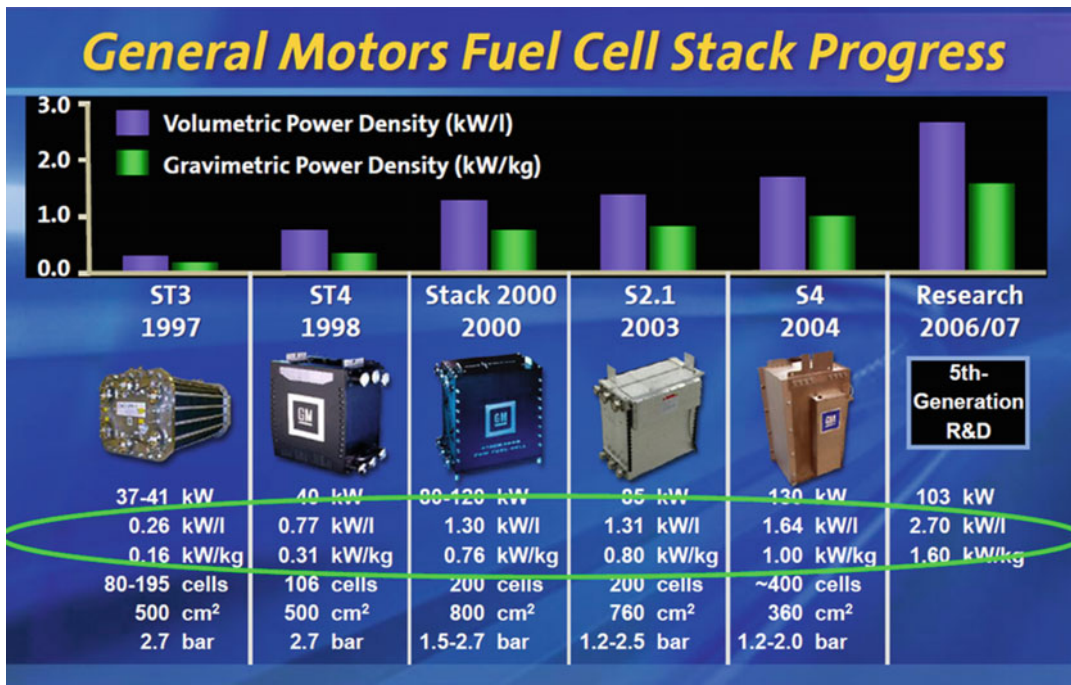
a



b

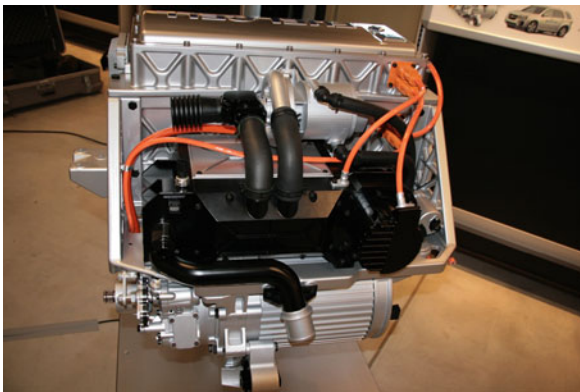
Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 3

(a) Timeline of fuel cell vehicle development at General Motors until 2001; (b) 2008 Chevy Equinox fuel cell vehicle [24, 25]



Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 4

Advancements in the development of the fuel cell stacks designed at General Motors from 1997 to 2004 is depicted [22]



Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 5

Photograph of fifth generation of General Motors PEMFC stack and system [23]

**UTC Power/UTC Fuel Cells** For the last 40 years, various groups under the umbrella of United Technologies (IFC, UTC Fuel Cells, UTC Power) have been uninterruptedly involved in the research and development of fuel cell stacks such as the commercial PAFC stack-power plant and the AFC PEMFC

stacks/power-plant used in the orbiter/space shuttle [26]. Over the last decade, UTC Power collaborated with several automotive companies to integrate their PEMFC stacks into various automotive platforms. Some of the interesting aspects of the UTC- PEMFC stack are that they operate close to ambient pressure and have unique water transport plates or separators; the plates are porous and have internal channels allowing for circulating water or coolant that performs the function of cooling the stack as well as passive water management. The reactants in the anode and cathode flow fields are, in principle, always humidified and improve the performance and durability of the membrane.

Some of the automotive companies they have worked with include: Hyundai-Kia Motor Company, Chevron Technology Ventures, Nissan, BMW, and a few others. UTC Power has also developed PEMFC 5 kW auxiliary power units APUs for BMW installed in BMW 7 series vehicles. The PEMFC stack APU provides energy for the vehicle's on-board electrical requirements. The third generation of the APU has been reported to perform for >3,000 h. UTC Power together with Hyundai-Kia Motor Company developed

an automotive FCV that was capable of starting and functioning under subzero conditions; they were tested in the winter of 2008 in Michigan under the US DOE Hydrogen Fleet and Infrastructure Program. UTC Power also supplied PEMFC stacks for initial generations of Nissan X-TRAIL FCVs.

**Ballard/AFCC** Ballard Power Systems was originally founded in 1979 as Ballard Research Inc. by Dr. Geoffrey Ballard to conduct research on high-energy lithium batteries. In the late 1980s and 1990s Ballard championed PEMFCs for automotive and other uses and re-invigorated the field. In 1995, Ballard Systems tested PEM cells in buses in Vancouver and Chicago and later in experimental vehicles made by DaimlerChrysler. In late 2007, Ballard pulled out of the hydrogen vehicle sector of its business to focus on fuel cells for forklifts and stationary electrical generation. Established in 2008, the Automotive Fuel Cell Cooperation (AFCC) is a Burnaby, B.C.-based joint-venture private company owned by Daimler AG, Ford Motor Company, and Ballard Power Systems Inc. to develop fuel cell stacks for automotive applications. Today, Daimler and Ford have more than 150 fuel cell vehicles on the road. The fuel cell vehicles of Ford and Daimler are combined together in this subsection since they both generally utilize stacks based on AFCC technology.

DaimlerChrysler unveiled a series of FCVs using Ballard stacks such as the NECAR 1 (50 kW, Compressed H<sub>2</sub>, 1994), NECAR 2 (50 kW, Compressed H<sub>2</sub>, 1996), NECAR 3 (50 kW, liquid methanol, 1997), NECAR 4 (70 kW, liquid H<sub>2</sub>, 1999), and F-Cell (A-class) FCV hybrid delivering 85 kW (Ballard Mark 900) using compressed H<sub>2</sub> in 2002. Most recently, Mercedes announced a series-production of B-class F-Cell powered by an electric motor that generates 136 hp and 214 lb-ft of torque, providing a range of 240 miles, and a refueling time of 3 min [27]. Ford Motor Company has also released a series of vehicles using Ballard stacks such as the P200HFC (1999), Focus FCV (2000), and Advanced Focus FCV (2002).

**Toyota** Toyota has been developing fuel cells since about 1996 with vehicles such as the RAV 4 FCEV-hybrid (Metal hydride, 1996), RAV 4 FCEV (Methanol, 1997), FCHV-3 (Metal hydride, 2001), FCHV-4 (Compressed H<sub>2</sub>, 2001), and FCHV-5 (reformed gasoline,

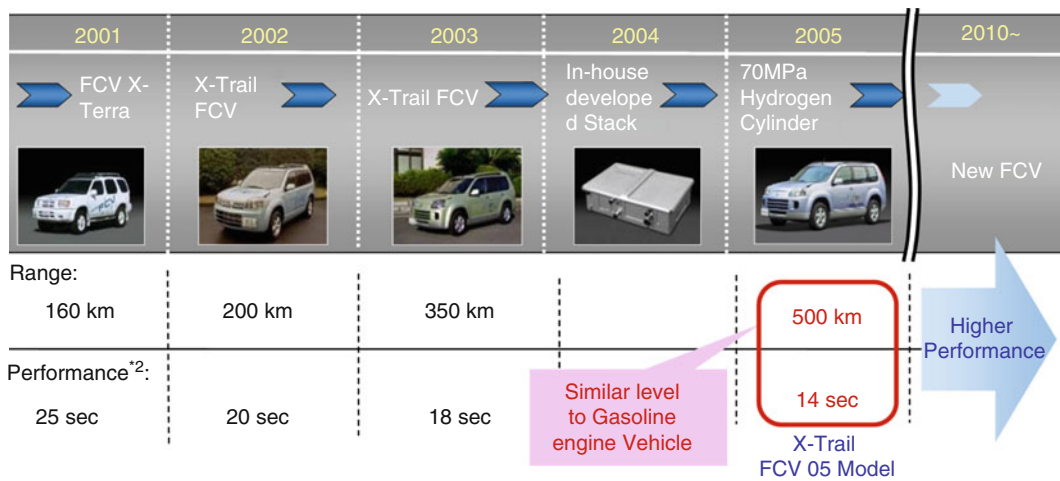
2001). Their efforts have resulted in the development of the latest Toyota Fuel-Cell Hybrid Vehicle (FCHV), some of which are being tested daily at the University of California in Davis and Irvine. Since 2001, a fleet of 25 FCHVs has accumulated more than 100,000 miles. In 2002, Toyota began limited marketing of the hydrogen-powered TOYOTA FCHV (fuel cell hybrid vehicle) in the USA and Japan [28]. Toyota FCHVs have also been successfully tested under subzero conditions. In real-world driving tests carried out in 2008 in collaboration with the US DOE, Savannah River National Laboratory, and National Renewable Energy Laboratory (NREL), the FCHV-adv averaged the equivalent of ~68 mpg achieving a range of ~430 miles on a single fill of compressed (70 MPa) hydrogen gas. Figure 6 depicts images of Toyota's FCHV and an under the hood look at their PEMFC stack [29].

**Nissan** Nissan's foray into the development of FCV technology started in 1996 in collaboration with various stack development partners such as Ballard and UTC Fuel Cells. In 1999, testing of a methanol-reforming fuel cell was initiated and in 2000 Nissan participated in the California Fuel-Cell Partnership (CaFP). Since 2004, Nissan has developed and used an in-house PEMFC stack in their FCVs. Their 2005 FCV employs 70 MPa hydrogen fuel tanks that allow the X-Trail FCV to have a driving range of about 300 miles. Figure 7 depicts the timeline for FCV development at Nissan along with metrics of driving range and acceleration time [30]. The range of their current generation vehicle is ~500 km (310 miles) and is comparable to a similar ICE powered vehicle. In 2007, they have developed a new generation of stacks that have significantly lower catalyst loadings, improved durability, high-rated power of 130 kW and a volumetric power density approaching 2 kW/L. Specifications of Nissan's 2005 model FCV are detailed in Fig. 8.

**Honda** A brief timeline for the development of FCVs by Honda is summarized below. Honda initially employed various 60–85 kW Ballard stacks in their initial FCVs such as the FCX-V1 (Metal hydride, 1999), FCX-V2 (Methanol, 1999), FCX-V3 (Compressed H<sub>2</sub>, 2000), FCX-V4 with ultra-capacitors (Compressed H<sub>2</sub>, 2001), and the FCX (Compressed H<sub>2</sub>, 2002) vehicle that was leased in Los Angeles.



**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 6**  
Images of Toyota's FCHV-adv installed with their PEMFC stack [29]



**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 7**  
Timeline for FCV development at Nissan along with metrics of driving range and 0–100 kmph acceleration time [30]

The US City of Los Angeles became the first FCX customer, leasing the first of five Honda FCX models for fleet use. In 2003, a next-generation fuel cell stack capable of power generation at temperatures as low as  $-20^{\circ}\text{C}$ , was announced. In 2005, Honda introduced the second generation FCX and the first to be powered by a Honda FC stack. FCX Clarity FCEV, a dedicated platform hydrogen fuel cell vehicle, debuted at the L.A. Auto Show; the new stack was 20% smaller and 30% lighter than its previous generation. Since 2008, FCX Clarity FCEV has been made available as a leased vehicle for consumer use [31].

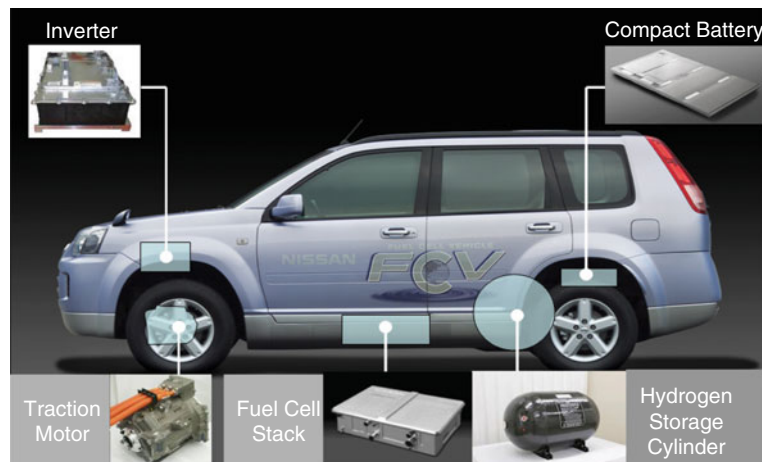
Some of the interesting aspects of the Honda fuel cell stack are the so-called V-Flow stack configuration

that claims easy drainage of water, higher cell stability, aromatic electrolytic membrane, improved thermal management, and improved packaging. Honda has also reported the use of specialized coatings with impregnated electrical contacts on their bipolar plates that reduced the contact resistance while being corrosion resistant. Figure 9a depicts the conventional stack assembly and Fig. 9b and c illustrate their improved efficient stack packaging design.

### Automotive PEMFCs

In this section, the operational modes that fuel cell stacks are subjected to; the 2015 performance, cost,

Classification		X-TRAIL FCV
Vehicle	Overall length / width / height (mm)	4485/1770/1745
	Curb Weight (kg)	1790(1860)
	Seating capacity	5
	Top speed (km/h)	150
	Cruising range (km)	Over 370 (over 500)
Motor	Type	<b>Co-axial motor integrated with reduction gear</b>
	Max. power (kw)	<b>90</b>
Fuel Cell Stack	Fuel cell	Polymer electrolyte type
	Max. power (kw)	<b>90</b>
	Supplier	In-house
Battery	Type	<b>Compact Lithium-ion Battery</b>
Fueling System	Fuel type	Compressed hydrogen gas
	Max. pressure (MPa)	35 (70)



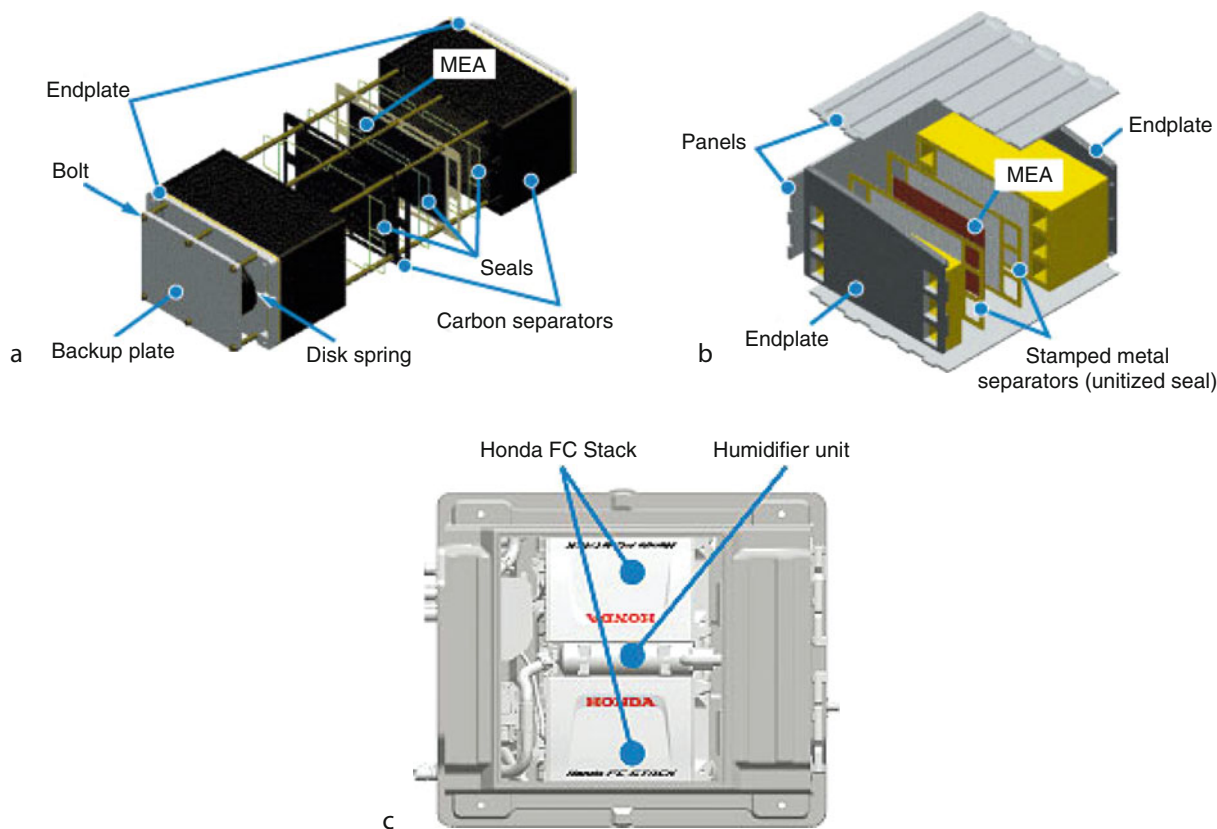
**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 8**  
 Detailed specifications of Nissan 2005 X-Trail FCV along with a layout of components of the fuel cell stack system [30]

and durability targets that fuel cell technology must achieve; and recent advancements made by automotive companies toward achieving the targets for commercialization are discussed.

**Operational Modes of Automotive PEMFCs**

Unlike residential and stationary fuel cells, automotive PEMFCs undergo the entire slew of aggressive variable

loads and environmental conditions that are typically experienced by conventional ICEs. The modes of operation of an automotive PEMFC can be simplified to the following [32]: (1) Idling/low load, (2) acceleration-deceleration/load cycling, (3) start-up shutdown, (4) cold temperatures/freeze-thaw cycles, and (5) contamination/impurities from the environments and cell degradation products. The impact of these modes of



**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 9**

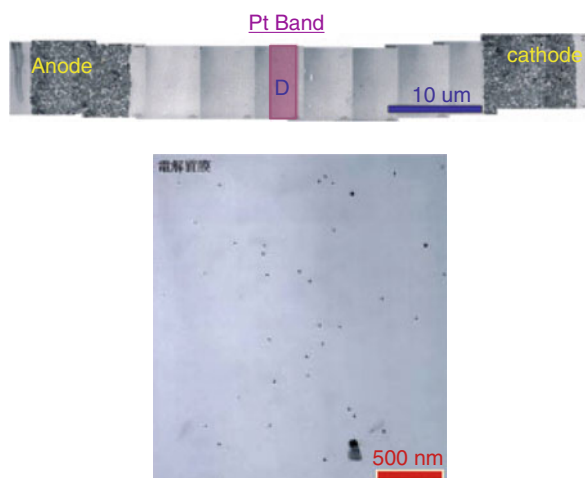
(a) Illustrates conventional stack package design of previous configurations and (b) and (c) illustrate Honda's new efficient stack package design [31]

operation on the PEMFC stack is briefly discussed below. Unless otherwise specified, the discussion assumes a Pt nanoparticle-based electrocatalyst on a high-surface-area carbon black support, a perfluorosulfonic acid membrane, hydrogen flow on the anode, air flow on the cathode with the reactants partially humidified at the inlet. Testing is performed on a variety of platforms such as liquid electrolyte-half-cells ( $\sim 0.198 \text{ cm}^2$ ), subscale fuel cells ( $25\text{--}50 \text{ cm}^2$ ), and full-size single cells or short stacks having active areas as high as  $400 \text{ cm}^2$ . The basic diagnostics techniques and characterization of fuel cell components can be found detailed in the literature [33–35].

**Idling/Low Load** When a vehicle in operation is at rest, for example, at a stop light, the current drawn by the fuel cell stack is low; essentially power is drawn by the auxiliary equipment in the vehicle. At such low current

densities, the cathode potential is high and may approach the open circuit voltage (OCV). Although the thermodynamic reversible potential at  $80^\circ\text{C}$  is about 1.18 V, leakage currents (few  $\text{mA}/\text{cm}^2$ ) especially due to hydrogen crossover across the ( $\sim 20\text{--}50 \mu\text{m}$  thick) membrane and electronic shorting lower the OCV to about 0.95 V. This high cathode potential leads to several degradation phenomena in the cells. The Pt-based catalyst dissolves at these potentials and diffuses toward the anode through the membrane and within 48 h forms a band inside the membrane as shown in Fig. 10 [36]. This Pt band formed in the membrane induces chemical degradation of the membrane itself due to formation of  $\text{OH}^\cdot$ , hydrogen peroxide, and peroxy radicals resulting in membrane thinning. The chemical degradation is accelerated under low RH conditions and higher temperatures. Degradation of the membrane leads to an increase in hydrogen crossover and can





### Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 10

Formation of a Pt band (with magnified view of the band) in the membrane when held under OCV conditions over a period of 100 h [36]

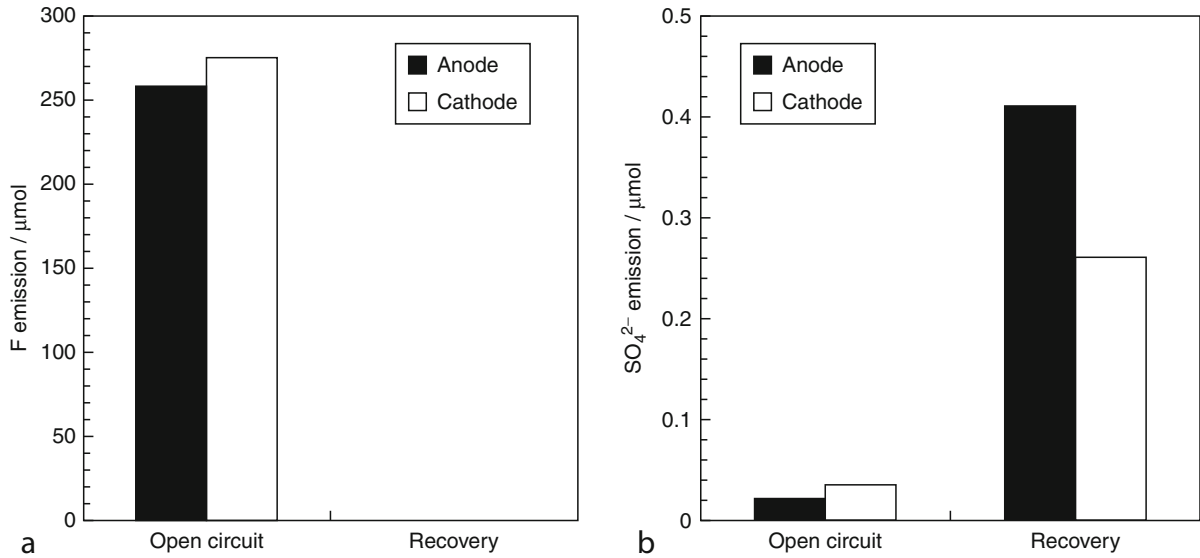
result in thinning, pinholes, and eventually catastrophic failure. Figure 11 shows the fluoride and sulfate elution from the exhaust water of a fuel cell during and after OCV testing [37]. About 6% of the fluorine in the membrane is released during 48 h of OCV hold. If the carbon support (on which the catalyst is dispersed) has a high surface area and therefore susceptible to corrosion, it may first become hydrophilic and over time corrode and oxidize leading to some Pt nanoparticle agglomeration which also leads to cell performance loss. If the fuel cell is a hybrid with a 10–20 kW NiMH or Li-ion battery, it may be possible to minimize the deleterious effects of high cathode potentials by using appropriate system controls and algorithms to lower the high potentials.

**Acceleration/Deceleration** PEMFC stacks in vehicles are subject to variable power demands that depend on environmental conditions such as the grade of the road as well as the driving behavior of the vehicle operator. As the load changes, the current drawn from the fuel cell and hence the cell voltage changes resulting in the cell being subject to load cycles or potential cycles. Thus, the fuel cell is subject to all kinds of potential cycles with the widest potential range being about 0.60–0.95 V [32, 35, 38–41]. The upper potential

corresponds roughly to OCV/idling and the lowest potential corresponds to the potential at peak power (about 0.6 V and a few amperes per square centimeter). The upper limit of the voltage is determined by the choice of the membrane (thickness/permeability) and in part by the design of the vehicle system controls. Although peak power is drawn from the stack for only ~5% of the time, it is necessary for the stack to have an electrical efficiency of ~55% (to be competitive with ICEs); also, the size of the stack is partly determined based on the peak power point.

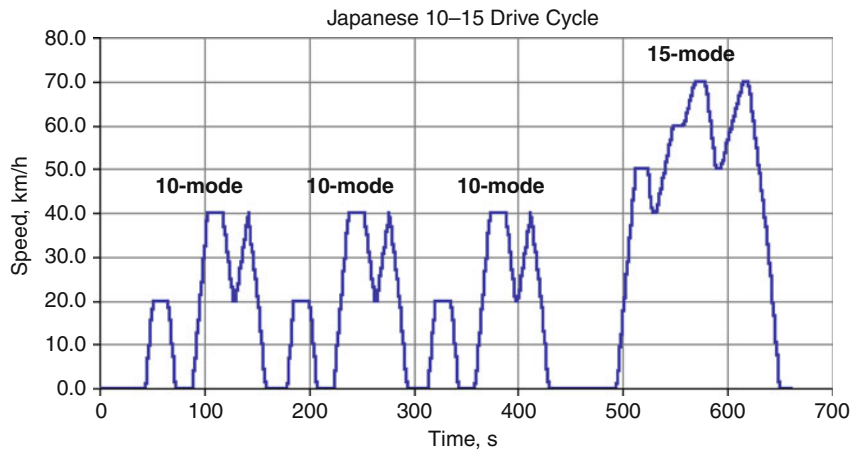
A number of automotive drive cycles (FUDS cycle, US06, NEDC, and Japanese 10–15 mode drive cycles, etc.) are available to represent the statistical usage of a vehicle and are used to determine the mileage of ICE vehicles. Figure 12 is an example of the Japanese 10–15 drive cycle and Fig. 13 an example of the US06 drive cycle. These drive cycles can be systematically converted into voltage-time profiles and applied to estimate the degradations rates of stacks under simulated driving conditions. A generic individual cycle profile is shown in Fig. 14 that represents a superset of all possible profiles [32]. The five elements of the profile that can be varied are: (1) duration at low potential, (2) duration of the ramp-up from low to high potential (or ramp-up rate), (3) duration at high potential, (4) duration of the ramp-down from high to low potential (or ramp-down rate), and, (5) duration at low potential. The high and low potentials can have infinitely different values between the upper and lower limit in the range 0.6–0.95 V.

It has been determined through a large number of studies that the widest potential cycles of 0.60–0.95 V cause the most degradation of the cathode catalyst layer. Automotives typically undergo 300,000 such cycles over their 5,000 h/10 years lifetime. At potentials above 0.95 V, the Pt surface is highly covered with oxide species while below 0.6 V it is almost free of oxide species as is observable from a typical cyclic voltammogram of platinum in acids or in a PEMFC. Thus the process of cycling in the range 0.60–0.95 V results in the growth (oxidation) and stripping (reduction) of oxide species on the surface and prevents the formation of a stable passivating film. (A more conservative potential regime of 0.70–0.90 corresponding to the iRFree potential that the catalyst layer experiences is sometimes selected.) Figure 15 depicts both the



Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 11

Fluoride emission during OCV hold test (6% of fluorine in the membrane is released over a 48 h hold) and sulfate emissions during recovery operation [37]



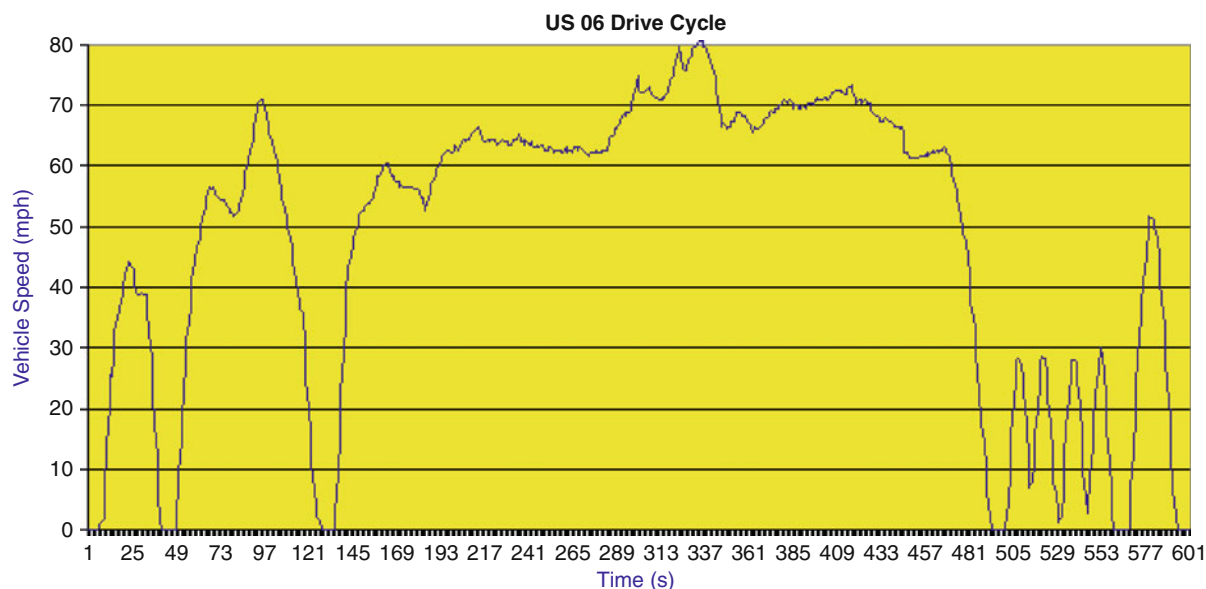
Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 12

Japanese 10-15 drive cycle for automobiles

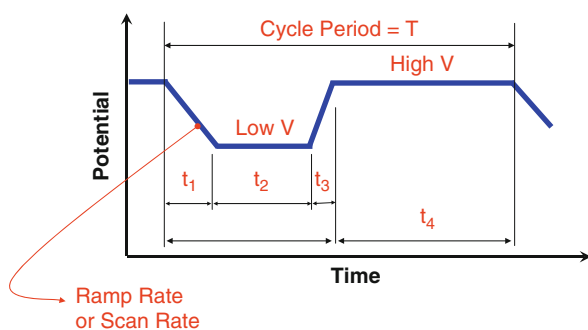
enhanced activity of PtCo/C electrocatalysts over Pt/C [42] and the surface area loss during cycling [39]. Based on recent research, it appears that exposure of bare (oxide-free) Pt to high potentials during the anodic portion of a scan in a cycle has a high impact on the Pt dissolution and degradation of the cathode catalyst layer. Again, practical methods that involve limiting the number of large cycles with the help of a battery in

a hybridized system mitigates the losses partially. Pt-based alloys, heat-treated Pt, and other modifications to the catalyst layer also provide partial material solutions that restrict the losses.

**Start-up/Shutdown** Automotive PEMFC stacks are shut down and started a number of times depending on the needs of the driver; the number of such occurrences is



Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 13  
US 06 drive cycle for automobiles



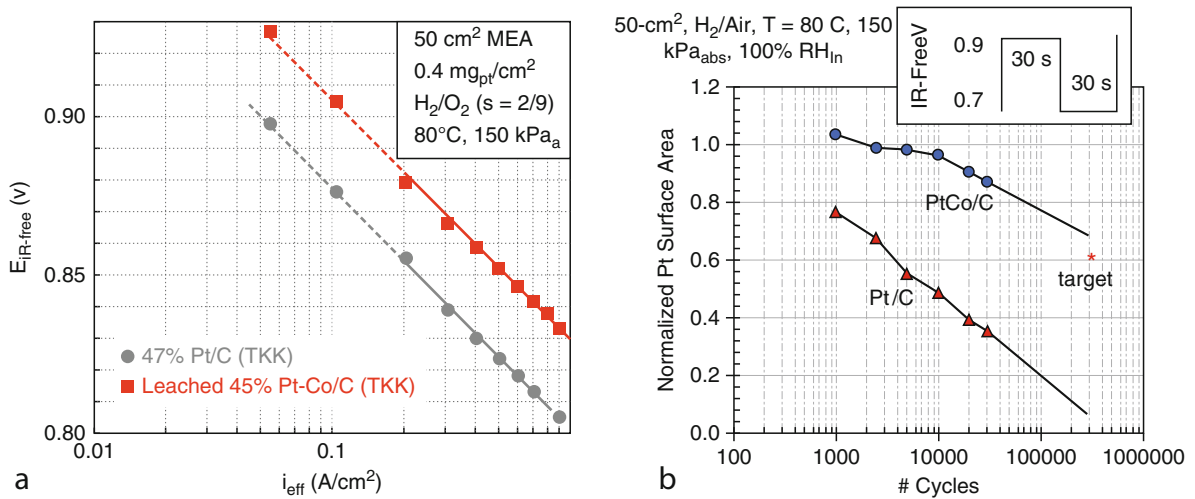
Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 14

Schematic representation of an arbitrary cycle extracted from a complex drive cycle illustrating the various parameters that can be varied in the cycle profile

about  $\sim 30,000$  over the stack life. When an automotive PEMFC stack is allowed to rest for a period of time, both the anode and cathode chambers get filled with ambient air from the atmosphere. When such a stack filled with air is restarted and hydrogen turned on to flow on the anode side, it pushes out the air leading to the formation of a "H<sub>2</sub>-air front." It was discovered around 2001 that frequent starting and shutting down of fuel cells resulted in a peculiar phenomenon where the cathode

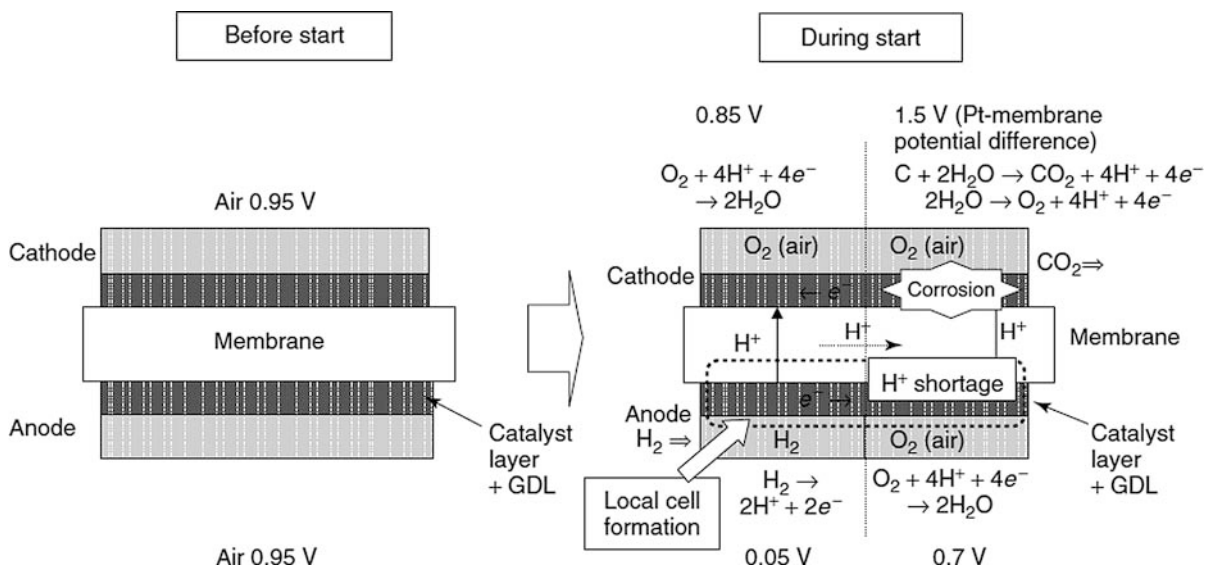
becomes subject to potential transients as high as 1.5 V [43–45]. At these high potentials, the carbon support of the cathode catalyst layer undergoes precipitous corrosion to carbon dioxide. Loss of the carbon leads to agglomeration of the Pt nanoparticles resulting in a loss in catalyst surface area, severe mass-transport issues, and a precipitous loss in cell performance. No losses are observed if the start-up takes place within a short time of shutdown as long as there is residual hydrogen in the anode of the fuel cell; thus the time interval between start-up and shutdown is an important parameter governing overall degradation rate [46].

Figure 16 below is a schematic of the start-up/shutdown phenomenon and the reactions that occur in the anode and cathode catalyst layers. Interestingly, no damage occurs on the anode-side catalyst layer. Figure 17 depicts the degraded cross section of an MEA (cathode facing up) after  $\sim 50$  start-up/shutdown cycles. Over the last decade, several mitigation techniques that involve controlling the procedure at start-up and shutdown have been reported and losses have been minimized, albeit at the expense of system complexity. Some of the mitigating solutions include: (1) shorting the stack to minimize the potential spike, (2) purging the anode using high flows of gases to minimize the H<sub>2</sub>-air front time (to less than 0.5 s),



**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 15**

(a) Enhanced activity ( $\times 3$ ) under H<sub>2</sub>/O<sub>2</sub> for pre-leached PtCo/C electrocatalysts compared to Pt/C as tested in subscale fuel cells [42]; (b) electrochemical area loss for Pt/C versus PtCo/C catalysts over 10,000 cycles in subscale fuel cells illustrates lower degradation for the PtCo/C [39]

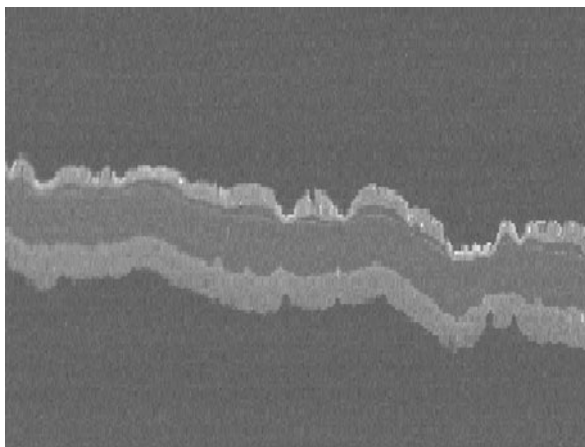


**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 16**

Schematic of start-up/shutdown phenomenon before and during passage of the H<sub>2</sub>-Air front through the anode chamber [43–46]

(3) drying the stack after shutdown, (4) lowering the Pt loading on the anode side to lower the ORR with a consequent suppression of the COR on the cathode, and (5) lengthening the time for which residual hydrogen remains in the anode, etc. The search for more

durable corrosion-resistant supports to replace carbon is underway so that the fuel cell system can be simplified. Oxides, nitrides, carbides, of tungsten, titanium, etc., are promising candidates but they often suffer from poor electronic conductivity, lower surface area



**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 17**

SEM cross section of an MEA with the cathode (top) layer showing severe corrosion and degradation after 50 uncontrolled start-up/shutdown cycles. A light band is also observed near the cathode-membrane interface due to the cycling of potential that takes place between 1 and 1.5 V

compared to carbon blacks and it can be difficult to disperse Pt nanoparticles on them [47].

**Cold Temperatures** Automotive PEMFCs must be capable of enduring cold weather conditions such as freezing temperatures as low as  $-40^{\circ}\text{C}$ . Even conventional ICE vehicles need help from engine block warmers, etc., in order to be able to start-up from subzero conditions. Under subzero conditions, an operating fuel cell will generate sufficient heat to prevent water from freezing and will likely show little or no loss in performance. The problem arises when a PEMFC has been idle for a long period of time under subzero conditions and is required to startup.

PEMFCs have the added challenge of liquid water in the system that can freeze, block channels and gas flow paths, and form icicles that may penetrate through the polymer membrane and cause damage. Blockage of channels can lead to fuel starvation, carbon corrosion, and degradation of the anode/cathode catalyst layer. The use of insulation, heaters, and antifreeze coolants are helpful for short periods of time, but the residual water within the fuel cell will eventually freeze forming ice. The catalyst layer and the diffusion media are especially susceptible to repeated freeze-thaw cycles

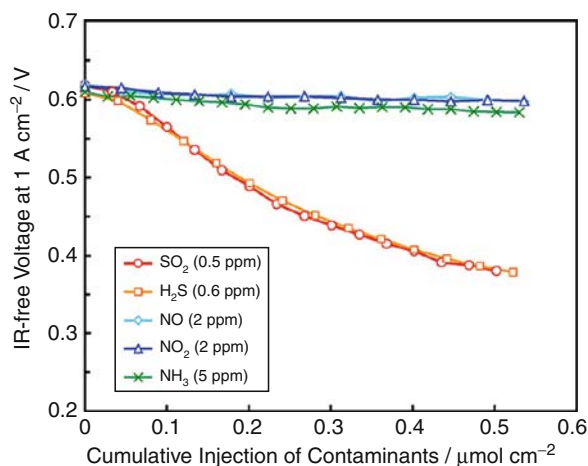
due to their porous nature and the inherent presence of water (loosely bound and free water) that can expand and contract leading to a disruption of the porous structure that is so important for mass-transport. Significant disconnection of the catalyst particles from the ionomer and carbon support could result in degradation in performance due to the loss in protonic and electronic interfacial contact.

Elimination or minimization of the water in the cells by purging with gases or vacuum drying prior to shutting down the stack is a general approach that has been found to be effective. In short, “keeping the stack warm” and “removal of water” are essentially the two main strategies to counter the cold start issue. The principal targets that automotive PEMFC stacks are required to meet are the ability to start unassisted from  $-40^{\circ}\text{C}$  and a cold start-up time to 50% of rated power in 30 s (from  $-20^{\circ}\text{C}$ ) with a start-up energy consumption of  $<5$  MJ [48]. Most automotive manufacturers have claimed to have achieved the targets although research to understand the phenomenon in detail continues.

**Contamination/Impurities** PEMFCs are very clean systems and act as filters for impurities introduced from ambient air, fuel used, and even degradation products from the cell materials. Both the membrane and the catalyst are susceptible to contamination and poisoning. Electrode degradation of PEMFCs can occur as a result of various impurities found in the fuel feed, air stream, as well as corrosion by-products from fuel cell components such as the bipolar plate, catalysts, or membrane.

Hydrogen for PEMFCs can come from various sources including reformed fossil fuels. Therefore, depending on the reforming technique and degree of posttreatment, small amounts of contaminants such as CO, CO<sub>2</sub>, NH<sub>3</sub>, H<sub>2</sub>S, etc., are expected to be present in the fuel stream [49]. Trade-offs in the level of impurities are unavoidable since ultrahigh purification would lead to elevated costs of hydrogen. The US FreedomCAR technology team has arrived at preliminary fuel mixture specifications that include:  $>99.9\%$  H<sub>2</sub>, 10 ppb H<sub>2</sub>S, 0.1 ppm CO, 5 ppm CO<sub>2</sub>, and 1 ppm NH<sub>3</sub>. Hydrocarbons such as methane, benzene, and toluene are other common impurity by-products from reforming processes.

Atmospheric air contains 78% nitrogen and 20% oxygen with the remainder being a number of trace gases and particulates depending on the local air quality. Pollutants found in the atmosphere include nitrogen oxides (NO and NO<sub>x</sub>), hydrocarbons, carbon monoxide (CO), ozone, sulfur dioxide (SO<sub>2</sub>), fine primary and secondary particulate matter, and chloride salts from the ocean and deicers. Chloride anions adsorb on Pt, occupy reaction sites, and significantly lower the ORR activity; fortuitously, the loss is recoverable simply by flushing out the anions with generated water. The degradation of performance due to 2.5 and 5 ppm SO<sub>2</sub> in the air stream was reported to be about 50% and 80%, respectively, by Mohtadi et al. [50]. The degradation is due to chemisorption of sulfur species on the Pt catalyst, and oxidation by the application of high potentials (CV) reversed the degradation but operation under normal potentials did not [51]. The poisoning mechanism by NO<sub>2</sub> was reported to be dependent on the time of exposure rather than bulk concentration and could apparently be reversed by operation under clean air for 24 h. Figure 18 depicts the loss in performance of a subscale fuel cell as a result of contamination of the catalyst in the presence of SO<sub>2</sub>, H<sub>2</sub>S, NO, NO<sub>2</sub>, and NH<sub>3</sub> [51]. The PEM is also easily contaminated by cationic impurities that may enter the stack through water, dissolved catalyst cations [52], as



**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 18**

The impact of several air contaminants on a fuel cell operated at 1 A/cm<sup>2</sup> [51]

well as corrosion products from the bipolar plate. The loss in cell performance is fairly severe due to the lowering of the conductivity of the membrane.

### The Status and Targets for Automotive PEMFC Commercialization

The overall status and targets for automotive PEMFC systems as well as hydrogen production, delivery and storage are outlined in Table 1. All of the targets need to be attained for successful commercialization of fuel cell vehicles.

The performance, durability, and cost of PEMFC system and stacks for automotives are approaching levels such that major automotive manufacturers have projected that commercialization will commence around 2015. The US DOE has had cost analysis conducted by independent organizations and estimates assuming 500,000 units/year indicate that system cost has been lowered from \$275/kW in 2002 to \$73/kW in 2008 [53, 54]. The target for PEMFC system cost in 2015 is \$30/kW. The durability of stacks under automotive conditions is currently around 2,500–3,000 h with the target in 2015 being 5,000 h.

In order to achieve commercialization in 2015, the targets of performance, durability, and cost of the fuel cell components (catalyst, membrane, diffusion media, bipolar plates, etc.) and the system (thermal, fuel, air, and water management, and balance of plant) have to

**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Table 1** Overall status and targets for PEMFC systems, hydrogen storage, production, and delivery [48]

	Status	Target
Fuel cell system cost	\$61/kW	\$30/kW
Fuel cell system durability	2,000–3,000 h	5,000 h
Hydrogen production	\$3–\$12/gge	\$2–\$3/gge
Hydrogen delivery	\$2.30–\$3.30/gge	\$1/gge
Hydrogen storage gravimetric	3.0–6.5 wt.%	7.5 wt.%
Hydrogen storage volumetric	15–50 g/L	70 g/L
Hydrogen storage cost	\$15–\$23/kWh	\$2/kWh

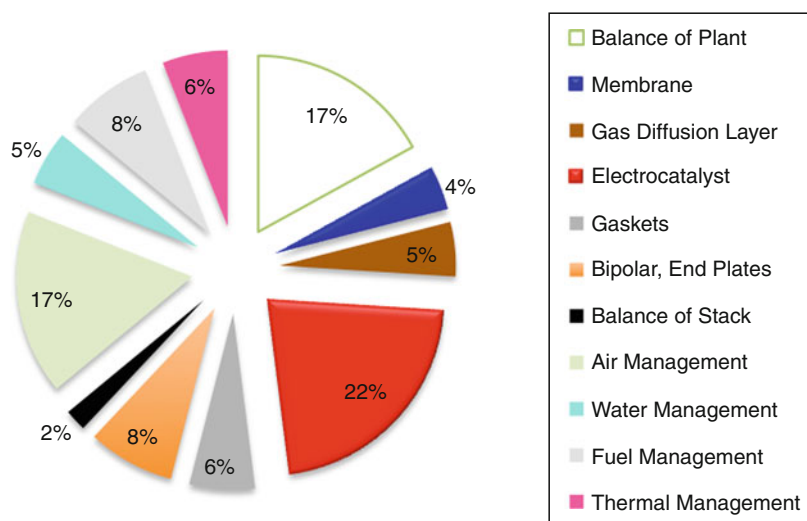
be met simultaneously. For example, high durability can easily be achieved by utilizing high loadings of Pt catalyst, corrosion-resistant bipolar plates with expensive coatings, etc., but this would violate the cost targets. It is required to maintain the performance of PEMFC stacks of today while at the same time lower the precious metal catalyst loadings by a factor of 4–10 and increase the stack durability by a factor of  $\sim 2$ .

**Basis for Targets** The targets of performance and durability of PEMFC stacks are loosely based on that of IC engines used in automobiles today. Durability targets are being addressed by organizations such as the US DOE's Office of Energy Efficiency and Renewable Energy's Hydrogen Fuel Cells and Infrastructure (HFCIT) program in close association with national laboratories, universities, and industry [48], the New Energy and Industrial technology Development Organization (NEDO) in Japan [55] and The European Hydrogen and Fuel Cell Technology Platform (HFP) and Implementation Panel (IP) in Europe. The Fuel Cell Commercialization Conference of Japan (FCCJ) have also published a booklet defining performance, durability, and cost targets in collaboration with Nissan, Toyota, and Honda and a summary has been reported elsewhere [56]. Overall, the general 2015 target for fuel cell stack durability (in cars) is roughly defined to be 5,000 h (150,000 driven miles) at the

end of which  $\sim 10\%$  performance degradation is allowable. This includes  $\sim 30,000$  cycles of start-up/shut-down,  $\sim 300,000$  cycles of wide span load, and hours of idling. So far  $\sim 3,000$  h of stack life with low degradation rates has been demonstrated and steady progress is being made toward the final goal. An additional durability enhancement by a factor of  $< 2\times$  is needed for attaining the commercialization targets of 2015.

The cost breakdown of PEMFC systems is as illustrated in Fig. 19: The most expensive component (22% of system cost) is the precious metal ( $\sim \$40/\text{g}_{\text{Pt}}$ ) catalyst used on the anode and cathode of each cell. Figure 20 shows the estimated reduction in cost of PEMFC stack system that has been achieved between 2002 and 2009. Figure 21 charts the breakdown of cost by subsystem and component for automotive PEMFC stacks.

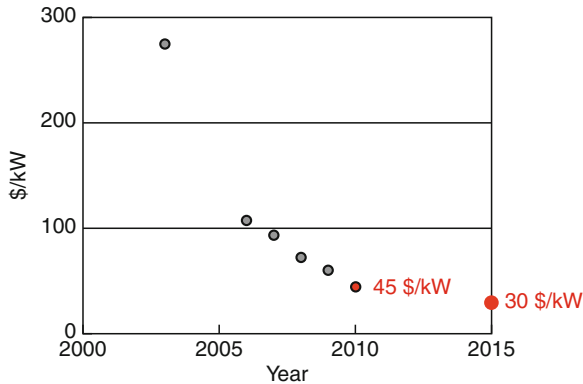
**Electrocatalyst Targets** The amount of Pt currently being used in ultra low emission IC engine automobiles (Pt, Pd, Rh, etc.) used in automotive catalytic converters can be as high as 6–10 g. This provides a justification for limiting the amount of Pt in a stack to  $\sim 10$  g (in a typical 100 kW stack) or  $\sim 0.1 \text{ g}_{\text{Pt}}/\text{kW}$ . (The target recommended by DOE is  $\sim 0.2 \text{ g}_{\text{Pt}}/\text{kW}$  [48]. Recycling of Pt used in fuel cells has been evaluated extensively by Tiax and the conclusion is that more than 90% of the Pt in fuel cell



**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 19**  
Cost structure for automotive PEMFC stack system [53, 54]

stacks is recyclable. Table 2 outlines the latest targets for automotive PEMFC electrocatalyst activity and costs.

**Membrane Targets** Today's membranes show excellent protonic conductivity ( $\sim 100$  mS/cm) at 100% RH and temperatures below  $80^\circ\text{C}$ . PFSA membranes have been

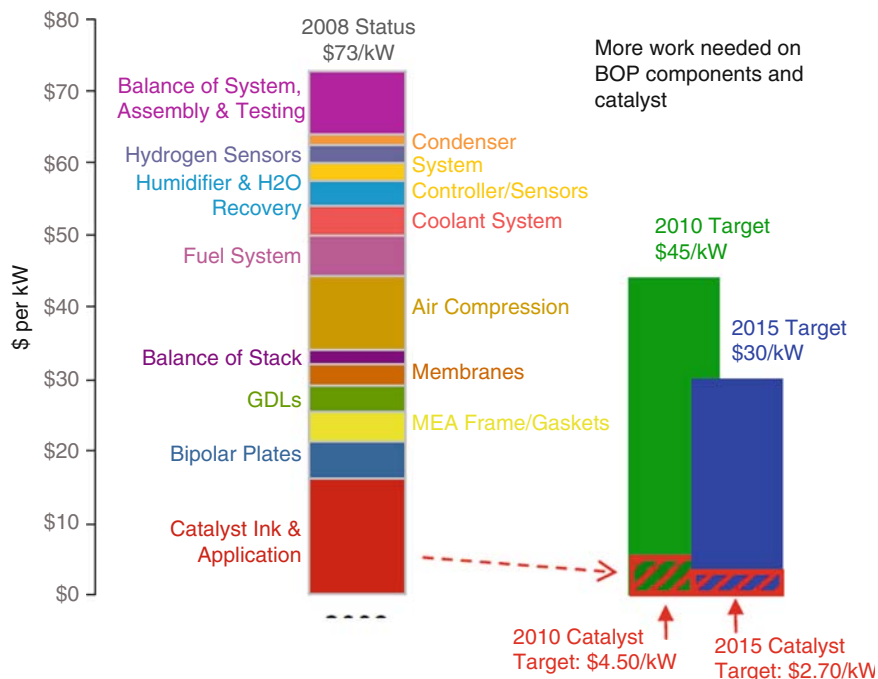


**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 20**

Reduction in cost of PEMFC stack system since 2002 and projected values and target for 2010 and 2015 [53]

improved over the last few years to also achieve durability values that are approaching the target values of 5,000 h. These improvements involve modification of PFSA membranes by varying the casting conditions, thermal pretreatments, forming composites, employing reinforcements, using different lengths of pendant side chains, fluorinating the end groups, etc. In particular, micro-reinforced membranes (such as those from W. L. Gore [57]) allow for the use of thinner membranes that result in the use of less ionomer, improve the back diffusion of water, lower the protonic resistance, and generally result in higher power density fuel cells.

For automotive operation, it would be helpful to have a membrane material that could function with high proton conductivity at a low RH and higher temperatures. It would allow for simplified humidification systems and improved heat rejection (smaller radiator) and improved cell performance (mass transport) is expected due to the absence of liquid water. Membranes that exhibit sufficient conductivity at high temperature by retaining water necessitate the use of higher stack pressures (and higher parasitic compressor-related



**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 21**

Breakdown of cost by subsystem and component for automotive PEMFC stacks [53]



**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications.** Table 2 Targets for automotive PEMFC electrocatalyst activity and costs. Mass Activity @ 0.90 V,  $\text{H}_2|\text{O}_2$ , 80°C,  $P_{\text{H}_2}$ ,  $P_{\text{O}_2}$  = 100 kPa (based on multiple sources)

	2010 Status	2015 Target
Mass activity	150 mA/mg	450–600 mA/mg
Mass of Pt per 100 kW stack	~30–40 g	~10 g

losses) especially as the temperature exceeds the boiling point of water at ambient pressure.

A cheaper hydrocarbon membrane is also desirable. Hydrocarbon membranes prepared from aromatic block copolymers constituted of alternating rigid sulfonic acid segments with hydrophobic polymeric units that are flexible has been reported. The unique morphology of the membrane provides comparable chemical stability as PFSA commercial membranes with high durability and a wide operating temperature range. Such membranes are already being used by automakers such as Honda. Other hydrocarbon-based membranes that are being actively researched include the use of disulfonated poly(arylene ether) sequenced block and random copolymers [58].

The key requirements for survivability and durability of membranes used in PEMFCs are mechanical durability (RH cycling, subzero start-up, shorting) and chemical durability (OCV holds). Survivability refers to the ability of the membrane to withstand operating conditions without the formation of pinholes that would lead to catastrophic hydrogen crossover and stack failure. Under subzero conditions, membranes can fail if the stack is not designed to eliminate most of the water in the flow fields to prevent formation of ice that can penetrate and damage the membrane on start-up. Durability refers to the slow thinning of the membrane over thousands of hours resulting in increased hydrogen crossover, membrane thinning, and consequent loss of cell performance. The membrane (restrained in a cell) is mechanically stressed (fatigue) when the fuel cell RH cycles between dry and wet conditions; thermal cycles also play a role. In the laboratory, tests have shown that non-reinforced PFSA membranes fail in a few hundred hours or ~6,000 cycles while reinforced membranes last as

long as 1,000 h or 60,000 cycles. Shorting is a mechanical degradation leading to electronic leakage currents through the membrane that may occur due to over-compression or penetrating irregularities from the catalyst layer or fibers of the diffusion medium. Shorting failures can be partially mitigated by using lower compression pressures (typically ~1,200 kPa), applying a coating of carbon black-based microporous layer (MPL) on the diffusion medium, etc. Chemical degradation of membranes is accelerated under OCV conditions where the potential is high and no water generated, as well as under generally low RH operation. Chemical degradation of the membrane is measured in terms of the fluoride release rate (FRR in  $\text{g F}^-/\text{cm}^2\cdot\text{h}$ ) in the collected water; the FRR exhibits Arrhenius dependence with temperature in the range 50–120°C. The FRR is of the order of  $1.0 \times 10^{-5}$  for typical membranes tested under OCV conditions and has been lowered with newer membranes to  $1.0 \times 10^{-7}$ . It should be noted that the mechanism of membrane degradation is not fully understood and is an area of intensive research. The two commonly invoked mechanisms are based on the assumption that hydrogen peroxide formed in the catalyst layer or at the Pt band in the membrane forms a hydroxyl radical which decomposes the membrane/ionomer by unzipping the less stable end groups or by scission of the main polymer chain. In actuality, mechanical and chemical degradation occur simultaneously in fuel cells accelerating the failure of the membrane.

The targets for membrane performance and durability based on figures reported by Nissan, USDOE, and the FCCJ are very similar. In order to be able to design a system that can reject heat at peak power, the fuel cell stack needs to operate for short spurts of time at temperatures as high as 90–95°C. At this time, most fuel cell stacks operate at about 80°C due to the restrictions of system complexity, parasitic (compressor) losses and humidification requirements. The membrane is targeted to operate at temperatures above 100°C (preferably 120°C) at RH <30% while exhibiting a membrane conductivity of >100 mS/cm. The durability targets are similar to that for the complete fuel cell and includes <10% performance loss over 5,000 h/10 years of operation under automotive conditions that include start-up/shutdown, load cycling, freeze (–40°C) and idling. The target cost of the membrane is \$10–20/m<sup>2</sup>

for production levels of 10 million m<sup>2</sup>/year. More detailed membrane performance targets base on USDOE and other sources are detailed in [Table 3](#).

**Bipolar Plate Targets** Graphite bipolar plates have been used for a long time and are still being used in some fuel cell stacks. Graphite has a very high electronic conductivity, low interfacial contact resistance (ICR) and high corrosion resistance. Disadvantages of graphite bipolar plates are the high cost of material and machining, increased stack volume, high gas permeability, etc., although high volume manufacturability is somewhat better with polymer–carbon fiber and composite bipolar plates. Graphite–polymer composite plates have also been researched to improve the manufacturability of graphite plates while maintaining conductivity and stability with temperature. Bare metal plates do not have the disadvantages mentioned above but suffer from high ICR and poor corrosion resistance that can lead to metal dissolution and possible poisoning of the PEMFC stack components. The dissolved metal ions can contaminate/lower the conductivity of the membrane by taking up ion exchange sites or adsorb on the catalyst and lower its activity. The stacks are generally tolerant to about 5–10 ppm of metal ion contamination. At the compression forces of relevance to PEMFCs, the USDOE bipolar plate areal resistivity target for ICR is 10 mΩ-cm<sup>2</sup>.

**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Table 3** Targets for selected properties of automotive PEMFC membranes [48]

	2015 Targets
Cost	\$5/kW; \$20/m <sup>2</sup>
Conductivity/resistance @ Operating T, RH	70 mS/cm; 20 mΩ-cm <sup>2</sup>
Conductivity/resistance @ -20 C	10 mS/cm; 200 mΩ-cm <sup>2</sup>
Min. electrical resistance	1.0 kΩ-cm <sup>2</sup>
Max. O <sub>2</sub> Crossover	<1 mA/cm <sup>2</sup>
Max. H <sub>2</sub> Crossover	<2 mA/cm <sup>2</sup>
Durability	>5,000 h
Survivability	-40°C to 120°C
RH cycles	20,000
OCV Lifetime	500 h

Stainless steels, as well as Al-, Ni-, and Ti-based alloys have been studied extensively as possible candidates for bipolar plates. One of the most well-studied materials for bipolar plates is SS 316/316L (16–18% Cr, 10–14% Ni, 2% Mo, rest Fe); other candidates are 310, 904L, 446, and 2205. Bare stainless steel plates form a passive 2–4 nm chromium oxide surface layer under PEMFC conditions that leads to unacceptably high ICRs. A similar trend is observed for the other alloys and therefore surface modification or surface coatings on selected substrate material has to be considered as a pathway to meet the technical targets of low ICR and high corrosion resistance.

Coatings may be metal based such as gold, TiN, CrN, metal carbides, etc.; carbon-based such as diamond-like carbon, graphite, graphite platelets, carbon–resin composites, conductive tin oxides, etc. Coatings have to be compatible with the substrate bare metal and have good adhesion and peel resistance.

An additional requirement that has to be satisfied is a high level of surface hydrophilicity. Non-wettable plate surfaces result in unstable reactant flows that have to be countered with higher pressure drops and gas stoichiometries and hence higher parasitic power drains. This requirement has led to the development of hybrid coatings that can provide the low ICR and high wettability. At this time, the bipolar plates account for 75% of the total stack weight and x% of the stack cost. [Table 4](#) summarizes some of the targets for bipolar plates for automotive fuel cells.

**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Table 4** Technical targets for automotive PEMFC bipolar plates [48]

	2015 USDOE Target
Cost	\$3.00/kW
Weight	0.4 kg/kW
Hydrogen permeation @ 80°C	2 × 10 <sup>-6</sup> cm <sup>3</sup> /cm <sup>2</sup> /s @ 80 C, 3 bar
Corrosion	1 μA/cm <sup>2</sup>
Resistivity	10 mΩ-cm
Flexural strength	25 MPa
Flexibility	3–5% deflection @ midspan

## Recent Progress in Fuel Cell Vehicles

Over the last decade, tremendous improvements have been made in PEMFC stacks used in fuel cell vehicles. Table 5 shows the status of selected metrics for the latest generation PEMFC stacks based on information published in the literature and through news announcements. Most of the FCVs produce peak power around 100 kW, use hydrogen compressed at 35 or 70 MPa (4–10 kg), and have a range, top speed, and acceleration that are approaching ICE vehicles. Most of these FCV are hybrids that employ a NiMH or Li-ion battery that helps improve the efficiency by storing energy during regenerative braking and deceleration and supplying energy during acceleration or idling.

In particular, some of the improvements reported by General Motors, Nissan, Toyota, and Honda in their latest generation stacks are discussed.

The main specifications for General Motor's Chevy Equinox are shown in Table 5; additionally, the vehicles employ a 93 kW generator, a 250 V, 35 kW Ni-MH battery and three carbon fiber pressurized (700 bar) hydrogen tanks. As part of a plan called "Project Drive-way," 100 of these vehicles were deployed in California, New York, and DC to gauge marketability based on customer response. Figure 22 shows schematics of the last two generations: GEN1FCs and GEN2FCS of GM's fuel cell systems comparing the improvements in various metrics. The Pt content in the GEN2FCS stack has been lowered to 30 g from

80 g and the stack and system volume reduced by almost 50%.

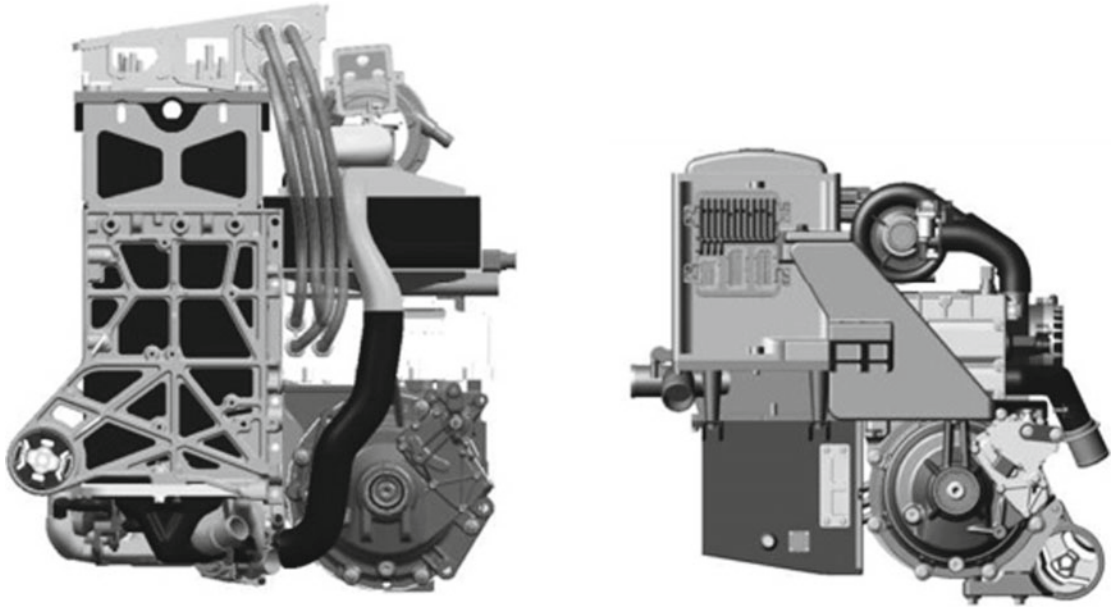
The specifications of the Nissan X-Trail FCV can be found in Table 5. Nissan has reported that their latest stacks incorporate reduced Pt catalyst loadings by a factor of 2, almost doubled the power density to 1.9 kW/L and improved the stack durability significantly. A comparison of their latest fuel cell stacks along with specifications is shown in Fig. 23.

Since the first Toyota FCHV launch in 2002, several improvements have been made to the technology. Toyota engineers have extended vehicle range and improved durability and efficiency through improvements in the fuel cell stack and the high-pressure hydrogen storage system, with additional cost reductions in materials and manufacturing. The FCHV-adv boasts a 150% improvement in range compared to the first-generation FCHV. In a real-world driving test carried out in 2008 in collaboration with the US Department of Energy, Savannah River National Laboratory and the National Renewable Energy Laboratory, the FCHV-adv averaged the equivalent of 68 mpg and achieved an estimated range of 431 miles on a single fill of hydrogen compressed gas.

The specifications for Honda's FCX Clarity that is being leased in Southern California can be found in Table 5; additional specs include a DC brushless motor of 100 kW and a 288 V Li-ion battery. Honda has claimed progress in several technologies that they have employed in their FCX Clarity FCV including

**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Table 5** Published metrics (multiple sources including manufacturer websites) for the latest generation of selected fuel cell vehicles

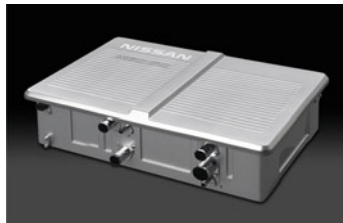
Vehicle/metric	Vehicle model	Power kW/hp	H <sub>2</sub> (kg, MPa)	Range (miles/km)	Time (0–60 mph) (s)	Top speed (mph)
GM	Chevy Equinox	104/140	4.2, 70	190/306	12	100
Nissan	X-Trail FCV	130/174	4.0, 70	310/500	10	95
Toyota	FCHV adv.	90/120	5.7, 70	520/835	10	97
Honda	FCX clarity	100/134	4.0, 35	240/386	9	100
Ford	Explorer FCEV	90/120	10, 70	350/564	18	87
Kia	Borrega FCV	115/154	7.0, 70	426/685	12.8	100
Daimler	F-cell, B-Klasse	90/120	4.1, 70	240/385	11.4	106



GEN1 FCS (GMT101X)		GEN2 FCS (GMT/E)
104	Stack Size (L)	64
80	Platinum (g)	30
405	System Size (L)	191
250	System Mass (kg)	130
62	Peak Power @ 150,000 miles (kW)	78
187	System Part Numbers	120
1903	System Part Count	1100

**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 22**

Improvements in power density, Pt loading, and system simplification for last two generations of General Motors FCV stacks [24]



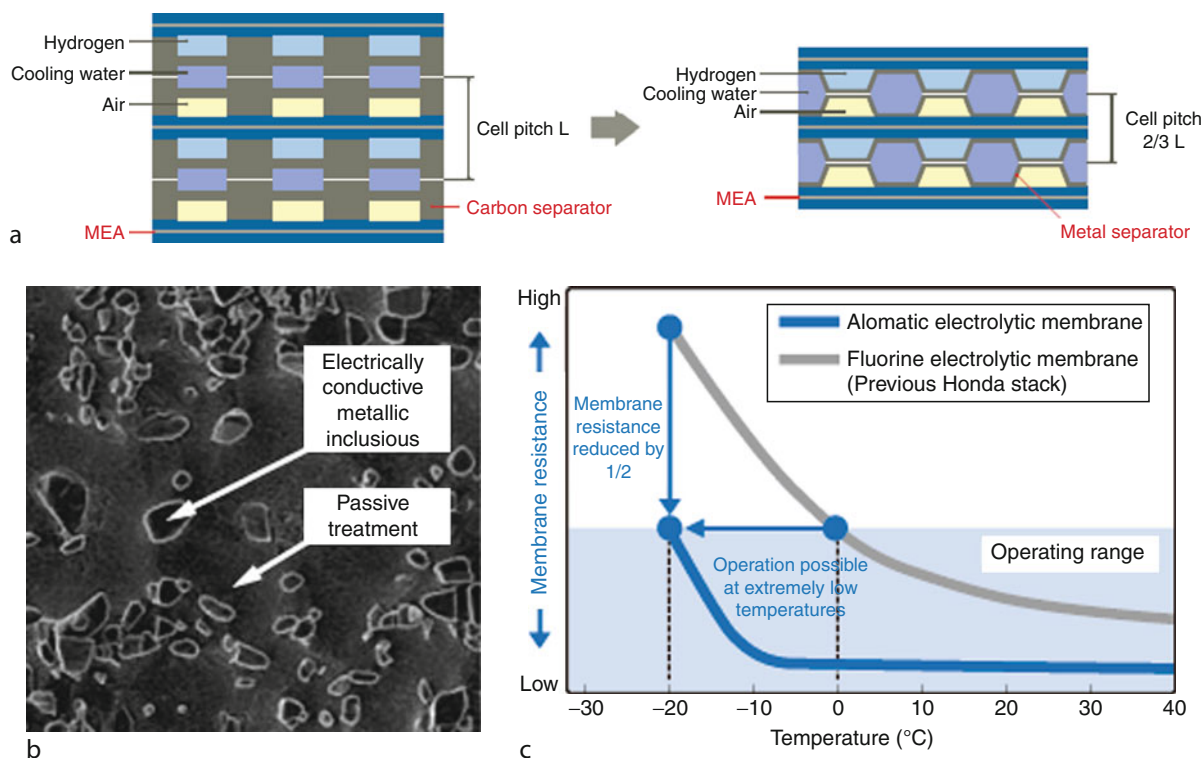
Specifications	
Max. Power	90 kW
Volume	90 L
Weight	116 kg



Specifications	
Max. Power	130 kW
Volume	68 L
Weight	86 kg

**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 23**

Improvements in stack power density for the last two generations of Nissan's in-house PEMFC stacks [46]



**Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 24**

Advances in Honda's recent automotive fuel cell stacks components of membrane and bipolar plates [31]

highly conductive aromatic membrane that allows the stack to be operated up to 90°C, thin, stamped metal plates having a high thermal and electrical conductivity. A two sub-stack (V-Flow stacks) configuration takes advantage of the stiffness of the metal separators to generate a structure in which the stack is encased by panels, resulting in efficient packaging. Figure 24 illustrates some of these improved technologies.

### Future Directions

Direct hydrogen-fuel-cell-powered vehicles have reached a level of development where the major automotive companies have publicly announced that initiation of commercialization is imminent around 2015. The targets of performance, durability, and cost agreed upon by various organizations, including the US DOE, appear to be achievable in the specified time frame. Well-delineated pathways and strategies have been established to address the barriers of cost and durability of PEMFC stacks and achieve the automotive targets.

The principal directions for reduction of cost and enhancement of durability of key fuel cell components, i.e., electrocatalysts, membranes, and bipolar plates are briefly summarized in this section.

Pt-based electrocatalysts will continue to be used in both the anode and cathode of automotive PEMFCs for the next decade. The current status of catalysts in automotive PEMFCs is a Pt-based carbon black-supported catalyst (20–50 wt.% Pt/C, 2–4 nm, 60–90 m<sup>2</sup>/g) with a loading of 0.20–0.35 mg<sub>Pt</sub>/cm<sup>2</sup> on the cathode and about 0.05 mg/cm<sup>2</sup> on the anode. About 30 g of Pt are required for a 100-kW-rated stack. The rated power usually corresponds to a voltage of 0.60 V at 1–2 A/cm<sup>2</sup>. The durability of fuel cell stacks under automotive conditions fall in the range of 3,500 h or 75% of the target life.

Progress has already been made over the years in enhancing the activity of the Pt nanoparticles dispersed on carbon support catalysts by alloying Pt with base metals such as Co, Ni, Fe, Rh, Cu, and Ti. The activity enhancements are a factor of 2–4, although the

performance decreases over time. The Pt-alloys/C also fortuitously exhibit improved durability in part due to their larger particle size and heat treatment. This is a conventional pathway that has already been explored and proven over the last few decades in PTFE-bonded electrodes used in PAFCs. A correlation to the improved activity from alloying is that the oxygen binding weakens as the d-band center shifts [59–61]; a volcano plot of various metals can be obtained with Pt sitting at the top [62]. Also, an increased number of neighboring metal atoms, 3d metal neighbors, as well as compressive strain have all been reported to weaken the oxygen binding to Pt. A positive shift in the onset of oxide observed in cyclic voltammograms indicating oxophobic nature of the surface has also been directly correlated to enhanced ORR activity. Larger Pt particles, heat-treated Pt particles, and alloys of Pt exhibit oxophobic tendencies and concomitant improved activities. More recently, the understanding of the impact of d-band center on the metal–oxygen bond strength has been refined [63]; the entire valence band structure (density of states vs binding energy) affects the bond strength and simplification of the band to a single “d-band center” is not valid. To summarize, the same binary and ternary alloys used previously in acid fuel cells are being implemented in PEMFCs using modified electrode structures.

A well-known weakness of the Pt-alloy approach is the leaching out of the base metal from the surface immediately and from the bulk over time. It has been shown by Stamenkovic et al. [64] that almost all the surface base metal is leached off once in contact with electrolyte; a Pt skeleton structure is formed (or Pt-skin for annealed catalysts) that has higher coordinated Pt atoms and a sublayer enriched in the base metal contributes to the higher activity. Nevertheless, as long as the catalyst activity and fuel cell performance is maintained within the target limits over the life of the stack, it remains an acceptable approach. Fundamental studies on single crystal Pt and Pt-alloys through a “materials by design” approach that involves the simultaneous application of a number of surface spectroscopies to a surface undergoing electrochemistry is being pursued by Markovic et al. [65]. Alloys with controlled crystal orientation such as the PtNi (111) have been demonstrated by Stamenkovic et al. [66] to show extremely high specific activities; it

remains a fundamentally important yet difficult approach to implement in practical catalysts. Perturbations in the approach to the use of alloy catalysts are the use of voltammetrically de-alloyed catalysts such as dealloyed Pt<sub>25</sub>Cu<sub>75</sub> and Pt<sub>20</sub>Cu<sub>20</sub>Co<sub>60</sub> nanoparticles [67]. Pt-rich surfaces and alloy-rich cores of such materials have been shown to exhibit significantly improved activity for oxygen reduction in acidic media due to a reduced Pt–Pt atomic distance (lattice strain) and also possess superior durability to commercial Pt/C.

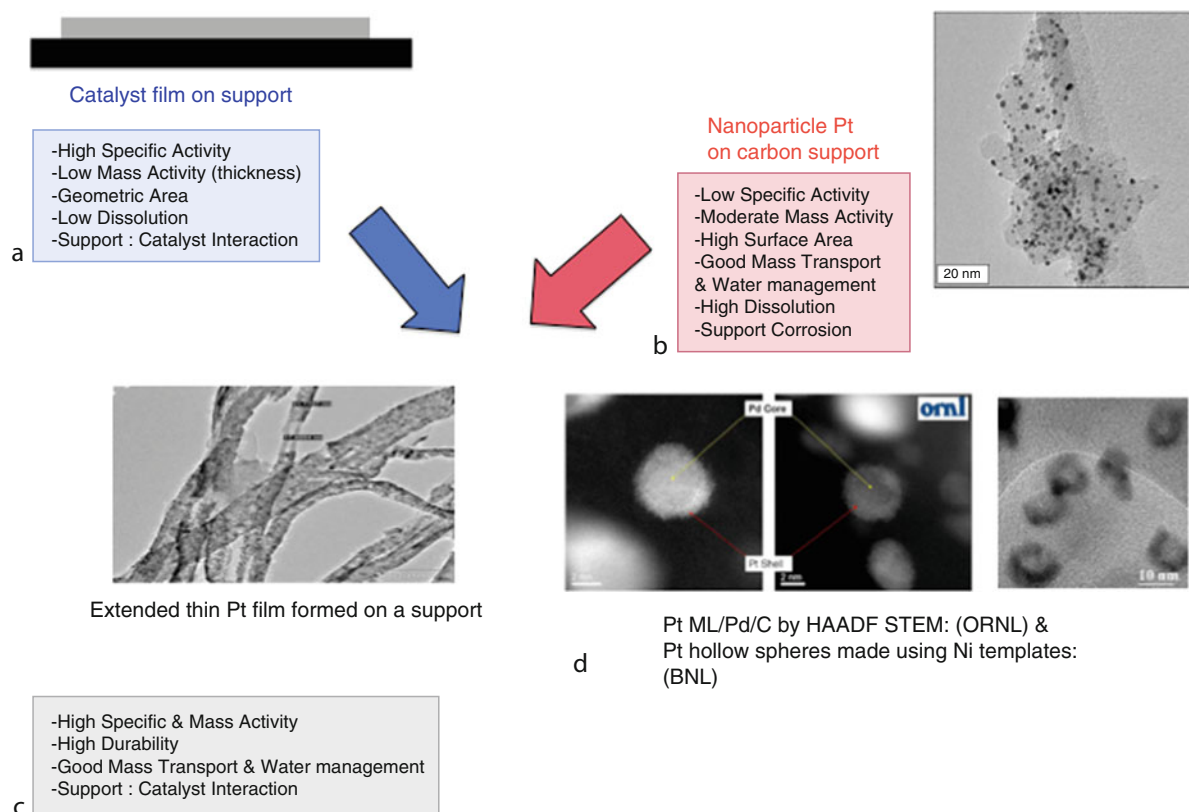
Two other pathways being explored to further drastically reduce the Pt loading are: (1) raising the catalyst mass activity (mA/mg) by using a core of base metal coated with monolayers of Pt and (2) applying the concept of high specific activity-extended thin films to a practical catalyst electrode system. The catalysts in the first pathway are often referred to as core-shell catalysts [68, 69]. A deposition technique that involves the replacement of a first UPD metal adlayer with a 2D deposit of a nobler metal monolayer to cover the surface uniformly forms the basis of the technique [70]. In addition to the use of expensive Pt only at the surface where reaction takes place, the use of different subsurface materials enhance the specific activity of the surface Pt layer; a Pt–metal mixed surface monolayer may also be used. About 4–10× enhancement in mass activity has been shown using these catalysts. The use of small quantities of gold clusters [71] on the surface has also shown to enhance the durability of the catalysts under potential cycling. Latest work in the area includes the generation of hollow Pt spheres (4–8 nm hollow spheres having 1–2 nm Pt shells) prepared from Ni templates exhibiting 5× enhancement of activity over solid nanoparticles (measured in rotating disk electrode liquid electrolyte half cells) [72]. Scale-up and evaluation in fuel cells of several of these technologies is currently underway.

For decades, discussion on the enhanced specific activity of larger nanoparticles has raged. Larger nanoparticles of Pt have been found to have both a higher activity as well as a higher durability by several groups [38, 73, 74]. In addition, bulk polycrystalline Pt has been shown to have ~10× higher specific activity than Pt nanoparticle in half-cells using ultra pure nonadsorbing liquid electrolytes such as perchloric acid. Over the last decade, the high specific activity of continuous thin films has finally been demonstrated in

practical MEAs of fuel cells especially by 3M with their NSTF catalysts [75, 76]. Two challenges remain in the widespread use of these catalysts in practical fuel cells: the mass specific surface area ( $\text{m}^2/\text{g}$ ) which is ten times lower for these films at this time and significant mass transport resistances due to flooding under humidified conditions due to the thinness of the cathode catalyst layer. Essentially, the catalyst films ( $\sim 30$  nm thick) are not thin enough to provide high mass activity and the electrode structure formed with these catalysts is not thick enough to disperse water easily. Both issues are being addressed by programs funded by the US DOE, and in laboratories elsewhere; the next 5 years will

determine the degree of success of this approach [77]. The potential for activity enhancement through these methods is predicted to approach a factor of 10; this would be sufficient to achieve the cost reduction target for catalysts used in the electrodes. Figure 25 outlines the main precious metal pathways for future fuel cell electrocatalysts.

Fundamental research related the formation of surface oxides on Pt [78–80] as well as Pt dissolution [81], although studied in the past, has come to the forefront again. This is due to the current understanding that the dissolution/surface area loss of Pt/C in automotive fuel cells is intensified by the cycling of load and hence



### Polymer Electrolyte Membrane (PEM) Fuel Cells, Automotive Applications. Figure 25

Pathways for future electrocatalyst development for automotive PEMFCs. (a) Thick films or bulk single crystal and polycrystalline catalysts that are ideal for fundamental studies on surface structure and mechanisms; these materials need to be modified into (c) and (d) to be applicable to fuel cells. (b) Typical commercial nanoparticles (2–4 nm) on a high-surface-area carbon support used in fuel cells at this time; (c) Thin continuous films of catalyst on a support such as carbon nanotubes that may provide a physical porous structure for mass transport in a fuel cell; (d) Core-shell catalysts where only the shell consists of precious metals and are supported on a typical high-surface-area support [72, 77, 89]

potential [35, 40, 82] along with the trend toward core-shell and nanofilms of catalysts. Operational methods to limit the losses through this understanding of mechanisms as well as material solutions are being actively pursued [32].

An intermediate approach to limit the use of Pt and instead use Pd-based metal alloys has shown reasonable success in activity improvement. Although fairly precious, Pd is currently four times less expensive than Pt. Pd alloyed with Mo, Ta, W, Re, and Cu have all been evaluated. A Pd–Cu (1:1) catalyst composed of 20 nm nanoparticles prepared by co-impregnation showed activity within range of Pt and is being further pursued [63, 83].

Non-precious metal catalysts (non-Pt group metals/non-PGM) research is still in its infancy and lags in activity and durability by a factor of  $\sim 10$  compared to conventional Pt/C electrocatalysts; they have shown significant progress in terms of improved activity in recently reported work, and research in this field continues as a long-term approach to completely eliminate platinum usage in fuel cells. The target for non-PGM catalyst activity is expressed per unit volume ( $>130 \text{ A/cm}^3$  @800 mV) since it is the volume/electrode thickness that determines its usability. An example is the ClFeN catalyst system that can be synthesized by heat treatment of Fe- $\text{N}_4$  macrocycles as well as individual precursors of the three elements [84, 85]. Lefevre et al. [86] synthesized the catalyst system by utilizing ball-milling to fill microporous carbon with iron ions and pore fillers to obtain higher ORR activity. Carbon supported as well as sputtered transition metal (Cr, Co, Fe)/chalcogens (Se, S) have also been studied with limited success [87]. In research funded by the USDOE, heat-treated, carbon-supported polypyrrole and PANI in the presence of salts of Fe and Co are being investigated in half-cell rotating disk electrodes as well as subscale fuel cells. In comparison to carbon, pyrolyzed carbon, PANI/C, PANI-Co/C, PANI-Fe/C, PANI-Fe<sub>3</sub>Co/C exhibited the highest ORR activity of  $27 \text{ A/cm}^3$  at 800 mV under  $\text{H}_2\text{O}_2$ ; the peroxide yield for PANI-Fe<sub>3</sub>Co/C was found to be the lowest at 0.5% [88]. A cyanamide-Fe-C ( $3.5 \text{ mg/cm}^2$ ) catalyst has shown catalyst activity in fuel cells of  $165 \text{ A/cm}^3$  @ 800 mV in most recent work [89]. Much work is remains to be carried out in the fundamental understanding and characterization of non-PGM catalysts in

terms of identification and quantification of the reaction sites and estimating the turnover frequencies and mechanisms for the ORR.

The high-surface-area carbon supports used today enable the use of small nanoparticles of Pt that are well-dispersed; the price paid in using a high-surface-area carbon black is that it tends to be susceptible to corrosion [90–93]. Carbon is thermodynamically susceptible to corrosion in the entire fuel cell operating regime but the kinetics are fortunately very slow. Near and above the open circuit potential of 0.95 V and especially above 1.1 V, the corrosion currents increase. Potentials in the range 1–1.5 V are seen only during uncontrolled start-up and shutdown of fuel cells [43, 94, 95]. Using mitigation techniques based on operating conditions, the highest potential seen is close to the OCV and the losses correspondingly lower. Nevertheless, some carbon corrosion is observed over the life of the fuel cell and material solutions are being actively sought. The material solutions will enable system simplification and lower costs. Alternative non-carbon supports such as titanium, tungsten oxides, nitrides, borides etc., are being studied that can provide a higher corrosion resistance at high potentials and also help anchor the Pt nanoparticles and limit degradation due to particle agglomeration. The main issues with alternative materials is that the corrosion resistance often comes with the cost of lower conductivity, lower surface area, different electrode structure, and the need for new methods to deposit catalysts on them.

Intensive research is being carried out to improve both the membrane performance (conductivity/areal resistance) and durability. The protonic conductivity is typically improved by the use of low EW ionomers; these ionomers swell with water to a greater extent increasing fatigue stresses. Thinner membranes are also being developed to reduce the areal resistance, but often need to be reinforced to provide mechanical strength and suffer from higher reactant permeability and susceptibility to shorting. Thus a trade-off between performance and durability is often unavoidable.

Development of new membrane materials that have the capability of sustaining proton conduction under low RH conditions and at temperatures as high as  $120^\circ\text{C}$  has been a struggle. Such a membrane would allow facile water management and also reduce thermal management issues in the stack. Of the two desired



properties, a membrane that operates at low RH with sufficient conductivity is more critical since catalyst (both platinum and carbon) degradation is also suppressed under these conditions.

The approach to synthesizing new membrane material candidates involves obtaining the required properties of conductivity, chemical and mechanical durability by: tailoring or modifying water retention, domain inter-connectivity, lowering the equivalent weight, use of additives such as metal composites and quenchers, stabilizing bonds and end group terminations, use of short side-chains, using reinforcements and cross-linking, raising the glass transition temperature, the use of amphoteric protogenic groups such as imidazole and phosphonic acid, etc. [96, 97]. Inorganics such as titania, silica, and zirconium phosphate have been incorporated into membranes in attempts to maintain conductivity at low RH through tightly bound water in these compounds. The use of amphoteric protogenic groups such as imidazole [98] and phosphonic acid that allow for proton transport in the absence of water has also been pursued [99]. PBI-PA-based membranes at temperatures above 200°C have been investigated by several groups [100] and commercialized by PEMEAS<sup>®</sup>. These materials suffer from the leaching of phosphoric acid, complex start-up and shutdown, as well as loss of Pt activity due to phosphate anion adsorption. Modifications to this approach include the use of sulfonimides, perfluorinated acids, metal oxides as additives to reduce the anion adsorption and increase the oxygen solubility. Heteropoly acids such as phosphotungstic acid, silicotungstic acid, sulfonated zeolites have been used with partial success [101]. 10–20 wt.% heteropoly acids (not immobilized) were combined with 3 M PFSA ionomers to cast membranes and prepare MEAs; they showed a reduction in the FRR in fuel cell tests under hot dry conditions [100]. Additives in general may not be stable and may leach out over long periods of time and may also change the mechanical properties of the membrane. High-temperature, water-free/water-insoluble, proton-conducting membranes (protic salt polymer membranes) where a salt repeat unit conducts protons with an adjacent unit without the transport of water are also being researched by Gervasio et al. [102]. The membranes are prepared from solvent-free liquid salts known as protic ionic

liquids or pILs that are tailored by selecting an acid and base.

Recently, the use of highly electron-poor poly(phenylene) backbones has resulted in ionomers containing sulfone ( $-\text{SO}_2-$ ) units connecting the sulfonic acid ( $-\text{SO}_3\text{H}$ ) functionalized phenyl rings that exhibit high proton conductivity and stability [103]. Some preparation routes have resulted in highly sulfonated material with an exchange capacity of  $\text{IEC} = 4.5 \text{ meq/g}$  or an EW of 220 g/eq along with low water transport coefficients. These materials though are water soluble and brittle in the dry state but may still be usable as a component in a PEM. The task of synthesizing and developing such membranes is a difficult task, but even partial success such as obtaining a membrane that can operate at similar temperatures as today but at lower RH will help advance PEMFC stack technology considerably. Hydrophobic-hydrophilic multiblock copolymers ( $\text{BPSH}_x\text{-BPS}_y$ ; where  $x = \text{MW of sulfonated poly(arylene ether)}$  and  $y = \text{MW of poly(arylene ether)}$ ) with varying block lengths and controlled morphology that develop order and produce a co-continuous hydrophilic phase for good conductivity at low humidity have also been the subject of research as potential candidates for a new PEM [58].

Incremental engineering modification of membranes is also being carried out with some success. Composite membranes that consist of conductive and nonconductive porous polymer reinforcements have been incorporated for some time into membranes to provide mechanical strength for extremely low thicknesses; they suffer from some loss of conductivity. One approach is to form a composite/polymer blend and decouple the proton conduction from other membrane requirements. Arkema has demonstrated an inexpensive hydrocarbon-based polyelectrolyte blended with polyvinylidene difluoride (PVDF or Kynar) that exhibits similar performance to commercial PFSA membranes [104]. Another approach involves the use of electrospun ionomer fibers embedded in a polymer [105]. 3 M has reported the fabrication of 825 EW non-reinforced membranes with new additives that meet the targets of 20,000 RH cycles and 500 h OCV test in fuel cells [106]. Giner Electrochemical Systems [107] has reported on the development of dimensionally stable membranes (DSM) with

laser-drilled supports composed of polysulfone or polyimide (Kapton) or that lowers the swelling of high acid content PFSA ionomers. Asahi Glass Co., [108, 109] has reported a new polymer composite membrane (based on PFSA) in PEMFCs operating at 120°C, 50% RH for ~4,000 h; the MEA tested in fuel cells had a degradation rate of 75  $\mu\text{V/h}$  and a FRR of less than 1% of the baseline control. Chemical modification of PFSA membranes is being carried out to minimize the non-fluorinated end groups susceptible to degradation. Additives such as  $\text{Ce}^{3+}$  and  $\text{Mn}^{2+}$  ions added in trace quantities into the membrane and ionomer have also been demonstrated to improve the chemical stability.

Although the focus of this article is on PEMs that are being used in automotive fuel cells today, brief mention must be made of work carried out on alkaline anion exchange membranes (AAEMs). Such membranes have the potential to exhibit sufficient activity when used with non-precious metal catalysts, may work with fuels such as methanol and ethylene glycol and provide some of the features that commercial PFSA membranes provide for PEMFCs. A cross-linked, water-insoluble,  $\text{OH}^-$ -conducting, alkaline polymer free of metal ions and consisting of counter ions bound to the quaternary-ammonium containing polymer backbone has been reported with reasonable preliminary results (133–153  $\mu\text{m}$  thick, 0.0092 S/cm @ 30°C, 100%RH) [110].  $\text{H}_2|\text{Air}$  alkaline membrane fuel cells that showed encouraging preliminary results with Pt and transition metal catalyst cathodes in  $\text{CO}_2$  free air have also been reported recently by Acta S.p.A. [111].

Graphite-based bipolar plates have been used in PEMFC stacks, they suffer from drawbacks such as higher manufacturing costs, greater thicknesses, higher gas permeability that is necessary for high power density automotive fuel cell stacks. Metal bipolar plates (stainless steel, Ni, Ti, Al-based alloys) on the other hand possess high thermal conductivity, high mechanical and flexural strength, and facile high volume production but tend to corrode and require corrosion-resistant and conductive coatings that increase their cost [112]. New coatings are under development along with thinner stamped plates with most of the details being proprietary at this time. An example is thermal nitriding of thin (0.1 mm foils, Fe-20Cr-4 V and type 2205) stainless steel plates to generate surface

layers of  $\text{Cr}_2\text{N}$ , CrN, TiN,  $\text{V}_2\text{N}$ , etc., that lowers the interfacial contact resistance and raises the corrosion resistance simultaneously [113]. The additional requirement of hydrophilicity to facilitate water management in the flow fields has led to the development of hybrid coatings that is capable of providing a low ICR and high wettability. Super-hydrophilicity has been shown by layer-by-layer deposition of silica nanoparticles onto bipolar plates, which meets the other requirements [114]. Electrostatic layer-by-layer techniques have been employed to generate 100 nm coating structures that are constituted from 5 to 10 nm graphite platelets and 19 nm silica nanospheres. A low ICR of 4  $\text{m}\Omega\text{-cm}^2$  and a high degree of hydrophilicity is simultaneously achieved by this method.

Figure 23 in the previous section showed the improvement in Nissans latest generation stack achieved by using metal separators instead of carbon. Figure 24 illustrates schematically novel surface treatments on stainless steel bipolar plates along with electrically conductive inclusions that help maintain high conductivity and corrosion resistance at the same time. Intensive applied research is being carried out to obtain thin, corrosion-resistant conductive bipolar plates modified with coatings that are inexpensive and conducive to high-speed/high-volume manufacturing.

A combination of synergistic improvements in the catalyst, support, gas diffusion layers, membrane, and essentially the entire porous electrode structure in conjunction with bipolar plates/flow fields is expected to improve the mass-transport of reactant gases, protons, and water management. Thus, an increase in the peak current density ( $\text{A/cm}^2$ ) and peak power density ( $\text{W/cm}^2$ ) will result; this in turn will lower the stack volume, the amount of catalyst, and membrane material used and raise the kW/L, kW/kg, and lower the \$/kW stack metrics. It should be noted that the rated or peak power for automotive stacks is based in part on maintaining an electrical efficiency of >50%; this dictates that the cell voltage has to be maintained above ~0.60 V. At this time, volumetric power densities of practical stacks in fuel cell vehicles have been reported to be as high as ~2 kW/L [46] and are likely to increase over the next few years contributing to lower stack costs (\$/kW).

Trends in short- and longer-term directions for key fuel cell components including electrocatalysts/supports,

membranes, and bipolar plates have been elaborated in this section; improvement of the performance and durability of these components will directly impact the entire automotive fuel cell system requirements, complexity, and cost. Durable catalysts with enhanced ORR activity, durable membranes that perform at very low humidity and durable bipolar plates that have low contact resistance will not only increase the power density and cost of the fuel cell stack but also simplify and lower/eliminate system component costs of the air compressor, humidification systems, recycle pumps, radiator, start-up/shutdown and freeze-start-related components, etc. A combination of advances in all the fuel cell components discussed above, system simplification, governmental policies that are sensitive to sustainable clean energy, and development of a hydrogen infrastructure will enable achieving the projected technical and cost targets needed for automotive fuel cell commercialization.

## Bibliography

### Primary Literature

- Grove WR (1842) On gaseous voltaic battery. *Phil Mag* 3:417
- Schonbein CF (1839) *Phil Mag* 14:43
- Mond L, Langer C (1889) *Proc Roy Soc* 46:296
- Jacques WW (1897) Electricity direct from coal. *Harpers Mag* 94:144–150
- Baur E, Tobler J (1933) Brennstoffketten Z *Elektrochem* 39:169–180
- Schmidt TJ, Paulus UA, Gasteiger HA, Behm RJ (2001) Peroxide rde, anion adsorption effect. *J Electroanal Chem* 508:41–47
- Tobler J (1933) *Z Elektrochem* 39:148
- Nernst W (1904) *Z Phys Chem* 47:52
- Tafel J, Emmert B (1905) Ueber die ursache der spontanen depression des kathodenpotentials bei der elektrolyse verduennter schwefelsaeure. *Zeit Phys Chem* 50:349–373
- Liebhavsky HA, Cairns EJ (1968) *Fuel cells and batteries*. Wiley, New York
- Vielstich W (1965) *Fuel cells*. Wiley-Interscience, London
- Maget HJR (1967) in: Berger C (ed) *Handbook of fuel cell technology*. Prentice-Hall, Englewood Cliffs, NJ, pp 425–491
- Liebhavsky HA, Grubb WT Jr (1961) The fuel cell in space. *ARS J* 31:1183–1190
- AFC A (2003) Alkaline fuel cells. In: Vielstich W, Lamm A, Gasteiger H (eds) *Handbook of fuel cells-fundamentals, technology and applications*. Wiley, New York
- Kordesch KV (1978) 25 years of fuel cell development (1951–1976). *J Electrochem Soc* 125:77 C–91 C
- Grubb WTJ (1959) US Patent 2,913,511
- Niedrach LW, Alford HR (1965) *J Electrochem Soc* 112:117
- Thomas CES (2007) Greenhouse gas results. [http://www.cleancaroptions.com/html/greenhouse\\_gas\\_results.html](http://www.cleancaroptions.com/html/greenhouse_gas_results.html)
- DOE US (2010) Well-to-wheels greenhouse gas emissions. [http://www.hydrogen.energy.gov/pdfs/9002\\_well-to-wheels\\_greenhouse\\_gas\\_emissions\\_petroleum\\_use.pdf](http://www.hydrogen.energy.gov/pdfs/9002_well-to-wheels_greenhouse_gas_emissions_petroleum_use.pdf)
- IPCC (2007) The IPCC assessment reports. <http://www.ipcc.ch/>
- Koppel T (1999) *Powering the future: The ballard fuel cell and the race to change the world*. Wiley, New York
- Taub A (2009) The opportunity in electric transportation. [http://www.ncsc.ncsu.edu/cleantransportation/docs/Events/2009\\_5-27\\_Taub\\_GM-EV.pdf](http://www.ncsc.ncsu.edu/cleantransportation/docs/Events/2009_5-27_Taub_GM-EV.pdf)
- GreenCarCongress: GM highlights engineering advances with second generation fuel cell system and fifth generation stack; poised for production around 2015. <http://www.greencarcongress.com/2009/09/gm-2gen-20090928.html>
- ChevyEquinox (2010) Chevy equinox fuel cell. <http://www.gm.com/vehicles/innovation/fuel-cells/>
- AutoBlogGreen (2010) 2008 chevy equinox fuel cell. <http://green.autoblog.com/photos/2008-chevrolet-equinox-fuel-cell/#380179>
- UTC Power (2010) UTC power: Transportation\automotive. [http://www.utcpower.com/fs/com/bin/fs\\_com\\_Page/0,11491,0151,00.html](http://www.utcpower.com/fs/com/bin/fs_com_Page/0,11491,0151,00.html)
- F-cell M-BB-c (2010) 2010 mercedes-benz b-class f-cell. [http://www.caranddriver.com/news/car/09q3/2010\\_mercedes-benz\\_b-class\\_f-cell-car\\_news](http://www.caranddriver.com/news/car/09q3/2010_mercedes-benz_b-class_f-cell-car_news)
- FCHV-adv T (2010) Fuel cell technology. <http://www2.toyota.co.jp/en/tech/environment/fchv/>
- Pressroom T (2010) Toyota fuel cell vehicle demonstration program expands. <http://pressroom.toyota.com/pr/tms/toyota/toyota-fuel-cell-vehicle-demonstration-151146.aspx>
- NissanHistory (2010) The history of Nissan's fuel-cell vehicle development. [http://www.nissan-global.com/EN/ENVIRONMENT/CAR/FUEL\\_BATTERY/DEVELOPMENT/FCV/index.html](http://www.nissan-global.com/EN/ENVIRONMENT/CAR/FUEL_BATTERY/DEVELOPMENT/FCV/index.html)
- Honda (2010) Honda: Fuel cell electric vehicle. <http://world.honda.com/FuelCell/>
- Uchimura M, Sugawara S, Suzuki Y, Zhang J, Kocha SS (2008) Electrocatalyst durability under simulated automotive drive cycles. *ECS Trans* 16:225–234
- Kocha SS (2003) Principles of mea preparation. In: Vielstich W, Lamm A, Gasteiger H (eds) *Handbook of fuel cells-fundamentals, technology and applications*. John Wiley & Sons, Ltd., New York, pp 538–565
- Kocha SS, Yang DJ, Yi JS (2006) Characterization of gas crossover and its implications in PEM fuel cells. *AIChE J* 52:1916–1925
- Uchimura M, Kocha S (2007) The impact of cycle profile on PEMFC durability. *ECS Trans* 11:1215–1226
- Ohma A, Suga S, Yamamoto S, Shinohara K (2007) Membrane degradation behavior during OCV hold test. *J Electrochem Soc* 154:B757–B760
- Sugawara S, Maruyama T, Nagahara Y, Kocha SS, Shinohara K, Tsujita K, Mitsushima S, Ota K-i (2009) Performance decay of proton-exchange membrane fuel cells under open circuit conditions induced by membrane decomposition. *J Power Sources* 187:324–331

38. Gasteiger HA, Kocha SS, Sompalli B, Wagner FT (2005) Activity benchmarks and requirements for Pt, Pt-alloy, and non-Pt oxygen reduction catalysts for PEMFCs. *Appl Catal B Environ* 56:9–35
39. Mathias MF, Makharia R, Gasteiger HA, Conley JJ, Fuller TJ, Gittleman CJ, Kocha SS, Miller DP, Mittelsteadt CK, Tao X, Yan SG, Yu PT (2005) Two fuel cell cars in every garage? *Electrochem Soc Interface* 14:24–35
40. Uchimura M, Kocha SS (2007) The impact of oxides on activity and durability of PEMFCs. *AIChE Journal Abstract No. 295b*
41. Uchimura M, Kocha SS (2008) The influence of Pt-oxide coverage on the ORR reaction order in PEMFCs. *ECS Meeting, Honolulu, HI 12–17 Oct 2008*
42. Kocha SS, Gasteiger HA (2004) The use of Pt-alloy catalyst for cathodes of PEMFCs to enhance performance and achieve automotive cost targets. *Fuel Cell Seminar, San Antonio, TX*
43. Reiser CA, Bregoli L, Patterson TW, Yi JS, Yang JD, Perry ML, Jarvi TD (2005) A reverse-current decay mechanism for fuel cells. *Electrochem Solid-State Lett* 8:A273–A276
44. Reiser CA, Yang D, Sawyer RD (2005) Procedure for shutting down a fuel cell system using air purge. *US Patent 6,858,336, 22 Feb 2005*
45. Reiser CA, Yang DJ, Sawyer RD (2005) Procedure for starting up a fuel cell system using a fuel purge. *US Patent 7,410,712, 12 Aug 2008*
46. Shimoi R, Aoyama T, Iiyama A (2009) Development of fuel cell stack durability based on actual vehicle test data: Current status and future work. *SAE International 2009-01-1014*
47. Merzougui B, Halalay IC, Carpenter MK, Swathirajan S (2006) Conductive matrices for fuel cell electrodes. *General Motors, US Patent Application 20060251954*
48. DOE US (2007) Fuel cell targets. <http://www1.eere.energy.gov/hydrogenandfuelcells/mypp>
49. Borup RL, Meyers JP, Pivovar B, Kim YS, Mukundan R, Garland N, Myers DJ, Wilson M, Garzon F, Wood DL, Zelenay P, More K, Stroh K, Zawodzinski TA, Boncella J, McGrath J, Inaba M, Miyatake K, Hori M, Ota K-i, Ogumi Z, Miyata S, Nishikata A, Siroma Z, Uchimoto Y, Yasuda K, Kimijima K-i, Iwashita N (2007) Scientific aspects of polymer electrolyte fuel cell durability and degradation. *Chem Rev* 107:3904–3951
50. Mohtadi R, Lee WK, Zee JWV (2004) SO<sub>2</sub> contamination. *J Power Sources* 138:216–225
51. Nagahara Y, Sugawara S, Shinohara K (2008) The impact of air contaminants on PEMFC performance and durability. *J Power Sources* 182:422–488
52. Kocha SS, Gasteiger HA (2004) Platinum alloy catalysts for PEMFCs. *Henry B. Gonzalez Convention Center, San Antonio, TX*. <http://www.fuelcellseminar.com/past-conferences/2004.aspx>
53. Satyapal S (2009) Hydrogen program overview. [http://www.hydrogen.energy.gov/pdfs/review09/program\\_overview\\_2009\\_amr.pdf](http://www.hydrogen.energy.gov/pdfs/review09/program_overview_2009_amr.pdf)
54. Satyapal S (2009) Fuel cell project kickoff. [http://www1.eere.energy.gov/hydrogenandfuelcells/pdfs/satyapal\\_doe\\_kickoff.pdf](http://www1.eere.energy.gov/hydrogenandfuelcells/pdfs/satyapal_doe_kickoff.pdf)
55. NEDO (2007) Nedo homepage. <http://www.nedo.go.jp/nenryo/gijyutsu/index.html>
56. Iiyama A, Shinohara K, Igushi S, Daimaru A (2009) Membrane and catalyst performance targets for automotive fuel cells. In: Vielstich W, Gasteiger HA, Yokokawa H (eds) *Handbook of fuel cells-advances in electrocatalysis, materials, diagnostics and durability*. John Wiley & Sons, Ltd.,
57. Cleghorn S, Griffith M, Liu W, Pires J, Kolde J (2007) Gore's development path to a commercial automotive membrane electrode assembly. <http://www.fuelcellseminar.com/past-conferences/2007.aspx>
58. McGrath J (2007) Advanced materials for proton exchange membranes. *DOE Hydrogen Program Merit Review Presentation*. [http://www.hydrogen.energy.gov/pdfs/review07/fc\\_23\\_m McGrath.pdf](http://www.hydrogen.energy.gov/pdfs/review07/fc_23_m McGrath.pdf)
59. Luczak FJ (1976) Determination of d-band occupancy in pure metals and supported catalysts by measurement of the L<sub>III</sub> x-ray absorption threshold. *J Catal* 43:376–379
60. Mukerjee S, Srinivasan S, Soriaga MP, McBreen J (1995) Role of structural and electronic properties of Pt and Pt alloys on electrocatalysis of oxygen reduction. *J Electrochem Soc* 142: 1409–1422
61. Nagy Z, You H (2002) Applications of surface x-ray scattering to electrochemistry problems. *Electrochim Acta* 47:3037–3055
62. Jalan V, Taylor EJ (1983) Importance of interatomic spacing in catalytic reduction of oxygen in phosphoric acid. *J Electrochem Soc* 130:2299–2302
63. Myers D (2009) [http://www.hydrogen.energy.gov/pdfs/review09/fc\\_20\\_myers.pdf](http://www.hydrogen.energy.gov/pdfs/review09/fc_20_myers.pdf)
64. Stamenkovic VR, Fowler B, Mun BS, Wang G, Ross PN, Lucas CA, Markovic NM (2007) Improved oxygen reduction activity on Pt<sub>3</sub>Ni(111) via increased surface site availability. *Science* 315:494–497
65. Markovic NM, Ross PN (2000) Electrocatalysts by design: From the tailored surface to a commercial catalyst. *Electrochim Acta* 45:4101–4115
66. Stamenkovic VR, Mun BS, Arenz M, Mayrhofer KJJ, Lucas CA, Wang G, Ross PN, Markovic NM (2007) Trends in electrocatalysis on extended and nanoscale Pt-bimetallic alloy surfaces. *Nat Mater* 6:241–247
67. Neyerlin KC, Srivastava R, Yu C, Strasser P (2009) Electrochemical activity and stability of dealloyed Pt-Cu and Pt-Cu-Co electrocatalysts for the oxygen reduction reaction. *J Power Sources* 186:261–267
68. Zhang J, Lima FHB, Shao MH, Sasaki K, Wang JX, Hanson J, Adzic RR (2005) Pt monolayer on noble metal-noble metal core-shell nanoparticle electrocatalysts for O<sub>2</sub> reduction. *J Phys Chem B* 109:22701–22704
69. Zhang J, Mo Y, Vukmirovic MB, Klie R, Sasaki K, Adzic RR (2004) Pt-Pd core-shell. *J Phys Chem B* 108:10955
70. Brankovic SR, Wang JX, Adzic RR (2001) Metal monolayer deposition by replacement of metal adlayers on electrode surfaces. *Surf Sci* 474:L173–L179

71. Zhang J, Sasaki R, Sutter E, Adzic RR (2007) Stabilization of platinum oxygen reduction electrocatalysts using gold clusters. *Science* 315:220–222
72. Adzic RR (2010) Contiguous platinum monolayer oxygen reduction electrocatalysts on high-stability-low-cost supports. [http://www.hydrogen.energy.gov/pdfs/review10/fc009\\_adzic\\_2010\\_o\\_web.pdf](http://www.hydrogen.energy.gov/pdfs/review10/fc009_adzic_2010_o_web.pdf)
73. Bregoli LJ (1978) The influence of platinum crystallite size on the electrochemical reduction of oxygen in phosphoric acid. *Electrochim Acta* 23:489–492
74. Makharia R, Kocha SS, Yu PT, Sweikart MA, Gu W, Wagner FT, Gasteiger HA (2006) Durable PEMFC electrode materials: Requirements and benchmarking methodologies. *ECS Trans* 1:3–18
75. Debe M (2005) Advanced meas for enhanced operating conditions, amenable to high volume manufacture. 2005 DOE Hydrogen Program Review. [http://www.hydrogen.energy.gov/pdfs/review05/fc3\\_debe.pdf](http://www.hydrogen.energy.gov/pdfs/review05/fc3_debe.pdf)
76. Debe M (2008) Advanced cathode catalysts and supports for pem fuel cells. [http://www.hydrogen.energy.gov/pdfs/review08/fc\\_1\\_debe.pdf](http://www.hydrogen.energy.gov/pdfs/review08/fc_1_debe.pdf)
77. Pivovar B (2010) Extended, continuous pt nanostructures in thick, dispersed electrodes. [http://www.hydrogen.energy.gov/pdfs/review10/fc007\\_pivovar\\_2010\\_o\\_web.pdf](http://www.hydrogen.energy.gov/pdfs/review10/fc007_pivovar_2010_o_web.pdf)
78. Conway BE (1995) Electrochemical oxide film formation at noble metals as a surface-chemical process. *Prog Surf Sci* 49:331–452
79. Conway BE, Barnett B, Angerstein-Kozłowska H (1990) A surface-electrochemical basis for the direct logarithmic growth law for initial stages of extension of anodic oxide films formed at noble metals. *J Chem Phys* 93: 8361–8373
80. Conway BE, Jerkiewicz G (1992) Surface orientation dependence of oxide film growth at platinum single crystals. *J Electroanal Chem* 339:123–146
81. Bindra P, Clouser SJ, Yeager E (1979) Pt dissolution in concentrated phosphoric acid. *J Electrochem Soc* 126:1631
82. Wang X, Kumar R, Myers DJ (2006) Effect of voltage on platinum dissolution relevance to polymer electrolyte fuel cells. *Electrochem Solid-State Lett* 9:A225–A227
83. Wang X, Kariuki N, Vaughney JT, Goodpastor J, Kumar R, Myers DJ (2008) Bi-metallic Pd-Cu oxygen reduction electrocatalysts. *J Electrochem Soc* 155:B602–B609
84. Jaouen F, Charretre F, Dodolet JP (2006) C-n4. *J Electrochem Soc* 153:A689
85. Medard C, Lefevre M, Dodolet JP, Jaouen F, Lindbergh G (2006) C-n4. *Electrochim Acta* 51:3202
86. Lefevre M, Proietti E, Jaouen F, Dodolet J-P (2009) Iron-based catalysts with improved oxygen reduction activity in polymer electrolyte fuel cells. *Science* 324:71
87. Campbell S (2005) Development of transition metal/chalcogen based cathode catalysts for PEM fuel cells. [http://www.hydrogen.energy.gov/pdfs/review05/fc13\\_campbell.pdf](http://www.hydrogen.energy.gov/pdfs/review05/fc13_campbell.pdf)
88. Zelenay P (2009) Advanced cathode catalysts. [http://www.hydrogen.energy.gov/pdfs/review09/fc\\_21\\_zelenay.pdf](http://www.hydrogen.energy.gov/pdfs/review09/fc_21_zelenay.pdf)
89. Zelenay P (2010) Advanced cathode catalysts. [http://www.hydrogen.energy.gov/pdfs/review10/fc005\\_zelenay\\_2010\\_o\\_web.pdf](http://www.hydrogen.energy.gov/pdfs/review10/fc005_zelenay_2010_o_web.pdf)
90. Bett JA, Kinoshita K, Stonehart P (1974) Crystallite growth of Pt dispersed on graphitized carbon black. *J Catal* 35:307–316
91. Bett JA, Kinoshita K, Stonehart P (1976) Crystallite growth of Pt dispersed on graphitized carbon black ii effect of liquid environment. *J Catal* 41:124–133
92. Cai M, Ruthkosky MS, Merzougui B, Swathirajan S, Balogh MP, Oh SH (2006) Investigation of thermal and electrochemical degradation of fuel cell catalysts. *J Power Sources* 160: 977–986
93. Kinoshita K (1988) Carbon electrochemical and physico-chemical properties. John Wiley & Sons, New York
94. Yu PT, Gu W, Makharia R, Wagner F, Gasteiger H (2006) The impact of carbon stability on PEM fuel cell start-up and shutdown voltage degradation. ECS 210th Meeting, Abstract 0598.pdf. <http://www.electrochem.org/meetings/scheduler/abstracts/210/0598.pdf>
95. Yu PT, Kocha SS, Paine L, Gu W, Wagner FT (2004) The effects of air purge on the degradation of PEMFCS during startup an shutdown procedures. Proc AIChE 2004 Annual Meeting, New Orleans, LA, April 25–29 (2004)
96. Kreuer KD, Paddison SJ, Spohr E, Schuster M (2004) PEM review. *Chem Rev* 104:4637–4678
97. Kreuer KD, Schuster M, Obliers B, Diat O, Traub U, Fuchs A, Klock U, Paddison SJ, Maier J (2008) Short-side-chain proton conducting perfluorosulfonic acid ionomers: Why they perform better in PEM fuel cells. *J Power Sources* 178:499–509
98. Schuster MFH, Meyer WH, Schuster M, Kreuer KD (2004) Toward a new type of anhydrous organic proton conductor based on immobilized imidazole. *Chem Mater* 16:329–337
99. Steininger H, Schuster M, Kreuer KD, Kaltbeitzel A, Bingol B, Meyer WH, Schauff S, Brunklaus G, Maier J, Spiess HW (2007) Intermediate temperature proton conductors for PEM fuel cells based on phosphonic acid as protogenic group: A progress report. *Phys Chem Chem Phys* 9: 1764–1773
100. Larson JM, Hamrock SJ, Haugen GM, Pham P, Lamanna WM, Moss AB (2007) Membranes based on basic polymers and perfluorinated acids for hotter and drier fuel cell operating conditions. *J Power Sources* 172:108–114
101. Meng FQ, Aieta NV, Dec SF, Horan JL, Williamson D, Frey MH, Pham P, Turner JA, Yandrasits MA, Hamrock SJ, Herring AM (2007) Structural and transport effects of doping perfluoro-sulfonic acid polymers with the heteropoly acids, h3pw12o40 or h4siw12o40. *Electrochim Acta* 53:1372–1378
102. Gervasio D (2010) Protic salt polymer membranes. [http://www.hydrogen.energy.gov/pdfs/review09/fc\\_06\\_gervasio.pdf](http://www.hydrogen.energy.gov/pdfs/review09/fc_06_gervasio.pdf)
103. de Araujo CC, Kreuer KD, Schuster M, Portale G, Mendil-Jakani H, Gebel G, Maier J (2009) Poly(p-phenylene sulfone)s with high ion exchange capacity: Ionomers with unique microstructural and transport features. *Phys Chem Chem Phys* 11:3305–3312

104. Yi J (2007) Development of low-cost, durable membrane and MEA for stationary and mobile fuel cell applications. [http://www.hydrogen.energy.gov/pdfs/review07/fc\\_9\\_yi.pdf](http://www.hydrogen.energy.gov/pdfs/review07/fc_9_yi.pdf)
105. Pintauro P (2010) Nanocapillary network proton conducting membranes for high temperature hydrogen/air fuel cells. [http://www.hydrogen.energy.gov/pdfs/review10/fc038\\_pintauro\\_2010\\_o\\_web.pdf](http://www.hydrogen.energy.gov/pdfs/review10/fc038_pintauro_2010_o_web.pdf)
106. Hamrock SJ (2010) Membranes and meas for dry, hot operating conditions. [http://www.hydrogen.energy.gov/pdfs/review10/fc034\\_hamrock\\_2010\\_o\\_web.pdf](http://www.hydrogen.energy.gov/pdfs/review10/fc034_hamrock_2010_o_web.pdf)
107. Mittelsteadt CK (2010) Dimensionally stable membranes. [http://www.hydrogen.energy.gov/pdfs/review10/fc036\\_mittelsteadt\\_2010\\_o\\_web.pdf](http://www.hydrogen.energy.gov/pdfs/review10/fc036_mittelsteadt_2010_o_web.pdf)
108. Endoh E (2008) Progress of highly durable mea for PEMFC under high temperature and low humidity conditions. *ECS Trans* 12:41–50
109. Endoh E, Terazono S, Widjaja H, Takimoto Y (2004) OCV degradation. *Electrochem Solid-State Lett* 7:A209–AA211
110. Varcoe JR, Slade RCT, Yee E (2006) An alkaline polymer electrochemical interface: A breakthrough in application of alkaline anion-exchange membranes in fuel cells. *Chem Commun* 1428–1429
111. Piana M, Boccia M, Filipi A, Flammia E, Miller HA, Orsini M, Salusti F, Santiccioli S, Ciardelli F, Pucci A (2010) H<sub>2</sub>/air alkaline membrane fuel cell performance and durability, using novel ionomer and non-Pt group metal cathode catalyst. *J Power Sources* 195:5875–5881
112. Wang H, Turner JA (2010) Reviewing metallic PEMFC bipolar plates. *Fuel Cells* 10:510–519
113. Brady MP, Wang H, Turner JA, Meyer HM, More KL, Tortorelli PF, McCarthy BD (2010) Pre-oxidized and nitrated stainless steel alloy foil for proton exchange membrane fuel cell bipolar plates: Part 1. Corrosion, interfacial contact resistance, and surface structure. *J Power Sources* 195: 5610–5618
114. Dadheech G, Elhamid MHA, Blunk R (2009) Nanostructured and self-assembled superhydrophilic bipolar plate coatings for fuel cell water management. *Nanotech Conference & Expo 2009*, vol 3, Technical Proceedings pp 18–183
- (2007) Scientific aspects of polymer electrolyte fuel cell durability and degradation. *Chem Rev* 107:3904–3951
- Conway BE (1952) *Electrochemical data*. Greenwood Press, Westport, CT
- Conway BE (1964) *Theory of principles of electrode processes*. Ronald Press, New York
- Conway BE (1995) *Electrochemical oxide film formation at noble metals as a surface-chemical process*. *Prog Surf Sci* 49:331–452
- Conway BE, Jerkiewicz G (1992) Surface orientation dependence of oxide film growth at platinum single crystals. *J Electroanal Chem* 339:123–146
- Gileadi E (1993) *Electrode kinetics*. VCH, New York
- Kinoshita K (1988) *Carbon electrochemical and physicochemical properties*. John Wiley & Sons, New York
- Kinoshita K (1992) *Electrochemical oxygen technology*. John Wiley & Sons, New York
- Kocha SS (2003) Principles of MEA preparation. In: Vielstich W, Lamm A, Gasteiger H (eds) *Handbook of fuel cells—fundamentals, technology and applications*. John Wiley & Sons, Ltd., New York, pp 538–565
- Kocha SS, Yang DJ, Yi JS (2006) Characterization of gas crossover and its implications in PEM fuel cells. *AIChE J* 52:1916–1925
- Koppel T (1999) *Powering the future: The ballard fuel cell and the race to change the world*. Wiley, New York
- Kreuer KD, Paddison SJ, Spohr E, Schuster M (2004) Pem review. *Chem Rev* 104:4637–4678
- Liebavsky HA, Cairns EJ (1968) *Fuel cells and batteries*. Wiley, New York
- Markovic NM, Ross PN (2000) Electrocatalysts by design: From the tailored surface to a commercial catalyst. *Electrochim Acta* 45: 4101–4115
- Mathias MF, Makharia R, Gasteiger HA, Conley JJ, Fuller TJ, Gittleman CJ, Kocha SS, Miller DP, Mittelsteadt CK, Tao X, Yan SG, Yu PT (2005) Two fuel cell cars in every garage? *Electrochem Soc Interface* 14:24–35
- Mench MM (2008) *Fuel cell engines*. John Wiley & Sons Inc., Hoboken, NJ
- Pourbaix M (1966) *Atlas of electrochemical equilibrium in aqueous solutions*, 1st edn. Pergamon Press, New York
- Prentice G (1991) *Electrochemical engineering principles*. Prentice Hall, Englewood, NJ
- Savodogo O (1998) Emerging membranes for the electrochemical systems: (i) solid polymer electrolyte membranes for fuel cell systems. *J New Mater Electrochem Syst* 47–66
- Vetter KJ (1963) A general thermodynamic theory of the potential of passive electrodes and its influence on passive corrosion. *J Electrochem Soc* 110:597–605
- Wilson MS, Gottesfeld S (1992) High performance catalyzed membranes of ultra-low Pt loadings for polymer electrolyte fuel cells. *J Electrochem Soc* 139:L28–L30
- Zawodzinski TA, Derouin CR, Radzinski S, Sherman RJ, Smith VT, Springer TE, Gottesfeld S (1993) Water uptake by and transport through nafion 117 membranes. *J Electrochem Soc* 140: 1041–1047

## Books and Reviews

- Alsabet M, Grden M, Jerkiewicz G (2006) Comprehensive study of the growth of thin oxide layers on pt electrodes under well-defined temperature, potential, and time conditions. *J Electroanal Chem* 589:120–127
- Bard AJ, Faulkner LR (1980) *Electrochemical methods*. John Wiley & Sons Inc., New York
- Bockris JO'M, Reddy AKN (1973) *Modern Electrochemistry: An Introduction to an Interdisciplinary Area*, Vol 1, Springer
- Borup RL, Meyers JP, Pivovar B, Kim YS, Mukundan R, Garland N, Myers DJ, Wilson M, Garzon F, Wood DL, Zelenay P, More K, Stroh K, Zawodzinski TA, Boncella J, McGrath J, Inaba M, Miyatake K, Hori M, Ota K-i, Ogumi Z, Miyata S, Nishikata A, Siroma Z, Uchimoto Y, Yasuda K, Kimijima K-i, Iwashita N

## Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction

ULRIKE I. KRAMM<sup>1</sup>, PETER BOGDANOFF<sup>2</sup>, SEBASTIAN FIECHTER<sup>2</sup>

<sup>1</sup>Brandenburg University of Technology Applied Physics and Sensors, Cottbus, Germany  
<sup>2</sup>Helmholtz-Zentrum Berlin für Materialien und Energie GmbH (HZB), Institut für Solare Brennstoffe und Energiespeichermaterialien, Berlin, Germany

### Article outline

Glossary

Definition of the Subject

Introduction

Transition Metal Carbides, Nitrides, and Chalcogenides

Non-noble Metal Catalysts with Molecular Centers

NNMC and Their Potential for PEM-FC Application

Future Direction

Bibliography

### Glossary

**R(R)DE** Rotating (Ring) Disk Electrode. This technique is an electrochemical standard method; in this work, the measurements are performed to determine the catalytic activity towards the oxygen reduction in liquid oxygen-saturated electrolytes. Depending on the rotation rate, the diffusion-limited current changes. Measurements at different rotation rates enable the calculation of the kinetic current by using the Koutecky-Levich equation.

**Fe-N-C catalyst** Group of catalysts for which it is believed that molecular FeN<sub>4</sub>- or FeN<sub>2+2</sub>-centers are responsible for the reduction of oxygen.

**Macrocyclic** Complex, organic molecule; in the context of this report, usually porphyrins, phthalocyanines, tetraazaannulenes, i.e., characterized by a tetrapyrrole core.

**NNMC** Non-noble metal catalysts: Catalysts prepared without any noble metals. As a result, the fabrication costs of such materials should be essentially lower as all basic-components are cheap.

**ORR** Oxygen Reduction Reaction. In this contribution, it stands for the electrochemical reduction of oxygen.

The favored pathway is the direct reduction to water whereas the indirect pathway via the formation of peroxides is undesirable for fuel cell application.

**Pyrochelates** Macrocycles which were heat-treated ( $T > 300^\circ\text{C}$ ). During pyrolysis, the initial molecular structure is transferred into a carbon matrix. As also some fractions of the molecular centers of the precursor are preserved, the resulting product is sometimes assigned as pyrocholate.

**Site density ( $S_D$ )** Number of active sites per volume of catalyst. For the calculation of site densities of Me-N-C catalysts, often two assumptions have to be made: (1) the density is similar to other carbon-based catalysts ( $0.4\text{ g/cm}^3$ ) and (2) each metal atom is related to an active site. In some cases, authors determine the exact mass density and/or number of active sites, so that the value becomes more accurate. In all other cases, it is usually overestimated as not all metal atoms are associated with active sites.

**Transition metal chalcogenides** A chemical compound which is composed of at least one chalcogen (O, S, Se, Te) anion and one or more transition metal cations.

**Turnover frequency (TOF)** The TOF gives the number of electrons which are transferred from the active site per site and second. Similar to the site density also here often assumptions have to be made.

### Definition of the Subject

More efficient energy conversion systems may help to reduce the use of fossil fuels and the emission of greenhouse gases. Polymer Electrolyte Membrane Fuel Cells are such devices which are of particular interest for automotive applications. Unfortunately, cost issues are still limiting the application of this technology in a highly competitive market. An important part of this is the inherently high cost of platinum which is commonly used as electrocatalyst. But recently, the replacement of platinum by NNMC has come into the focus of reach. The most promising approaches – comprising the use of crystalline phases and catalysts with molecular active centers – are described in this entry; limitations of both classes of materials are discussed.

### Introduction

Although platinum and platinum alloys are the state-of-the-art electrocatalysts for PEM fuel cell applications,

the platinum loading required to reduce the effective overpotential of the oxygen reduction reaction at high current densities ( $1\text{--}2\text{ A/cm}^2$ ) to an acceptable level is still quite high for standard electrode structures ( $0.2\text{--}0.5\text{ mg/cm}^2$ ). Therefore, there have been many attempts to improve the dispersion of the nano-scaled platinum particles and to optimize the electronic states of the catalytically active platinum interface with the aim to increase the mass-related activities of the platinum catalysts.

Pt-alloys PtMe (Me = Cr, Mn, Co, Ni, V, Ti) were investigated early on, since they show an increased specific activity with respect to ORR [1, 2]. Enrichments and depletions of alloying metals on the particle interface were found, which influenced the electrochemical activity significantly. Predominant scientific success was achieved in the field of “Pt-monolayer catalysts” [3], the concept of “Pt-skin electrocatalysts” [4], and the concept of “unalloyed Pt bimetallic catalysts” [5].

Some authors report a rise in the mass specific activity of these catalysts by a factor up to 20 in PEM fuel cells [6]. Considering the complex production methods, it remains a question to what extent this concept will effect a real cost reduction. Furthermore, it has to be proven to which extent these sub-nanometer-structured catalysts remain stable in an operating fuel cell under corrosive conditions. Commercial platinum catalysts already show a significant reduction of the active platinum surface especially under changing loading conditions during the long-term operation of a fuel cell due to dissolution and aggregation of Pt particles, which leads to a decline of the fuel cell efficiency [7, 8].

Further challenges remain unsolved concerning the ability of platinum to catalyze the oxidation of the carbon support and the poisoning of the platinum interface in the presence of traces of toxic gases such as CO or H<sub>2</sub>S, present in the environment, which arrive at the fuel cell via the cathodic air flow. It needs to be mentioned that novel research activities aim to operate the PEM fuel cells at higher temperatures ( $>100^\circ\text{C}$ , HT-PEM-FC) in order to obtain enhanced reaction kinetic and, therefore, higher efficiencies. However, due to the elevated temperature, the dissolution of platinum as well as the platinum-catalyzed carbon oxidation both will increase.

In spite of the success in the optimization of platinum catalysts, a major breakthrough in the field of fuel

cells is yet to be achieved. Especially, the desire for a significant cost reduction by the replacement of platinum motivates international research activities investigating new catalyst concepts for cathodes. Thereby, the cathodes have to be sufficiently stable under fuel cell conditions; the alternative non-noble metal catalysts (NNMC) need to have a high selectivity for direct reduction of oxygen to water. The US department of energy (DOE) defined 25% of the achievable current density of a commercial platinum catalyst as target value for 2015. For fuel cell application, the catalysts should be producible in such a nano-structured form that suitable gas diffusion structures can be built.

Even if the numerous investigated non-noble material systems cannot yet fulfill all necessary demands, they do highlight new ways and principles in the electroreduction of oxygen, which could become applicable after a further aim-oriented development of these groups of materials. In the next sections, a selection of NNMC will be presented, which constitute a promising potential due to their catalytic properties.

### Transition Metal Carbides, Nitrides, and Chalcogenides

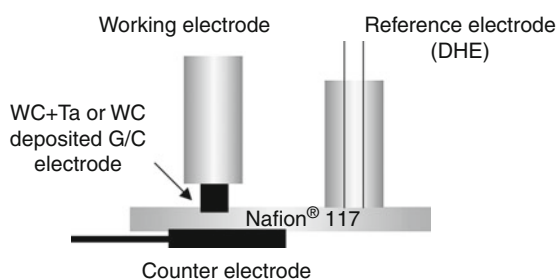
As reported so far, one of the best platinum-free ORR catalysts of chalcogenide-type structure is a selenium-modified ruthenium catalyst (RuSe<sub>x</sub>/C) [9–20]. State-of-the-art catalysts are composed of carbon-supported nano-scaled ruthenium particles whose surface was modified with selenium [9–14]. The modification leads to 10 times higher ORR activity, protects the ruthenium particles against electrooxidation, and suppresses the H<sub>2</sub>O<sub>2</sub> formation. As RuSe<sub>x</sub>/C is insensitive to methanol, it might be particularly suitable as an alternative cathode material in direct methanol fuel cells (DMFC) where platinum shows potential losses due to the methanol crossover [15–18]. However, ruthenium is still a costly and rare noble metal and seems not to be a feasible alternative to platinum. Therefore, readers who are interested in this type of catalyst are referred to the cited literature.

Because of a high electrical conductivity, transition metal carbides and nitrides are promising materials for electrocatalysis. Tungsten carbide has proven to possess platinum-like characteristics in terms of the valence band structure at the Fermi level and the chemisorption of



oxygen and hydrogen [21–24]. Therefore, it was intensively investigated for the electrooxidation of hydrogen and methanol [25–28] as well as for the electroreduction of oxygen. Mazza and Trasatti [29] studied several transition metal carbides (TMe: Ti, Ta, W) in sulfuric acid and selected WC as the most active material for the ORR among the others. However, tungsten carbide revealed poor corrosion stability in acidic electrolytes. More recent investigations on WC by Lee et al. [30] demonstrated that the stability and activity of these catalysts can significantly be increased by the addition of tantalum. They prepared pure WC and WC + Ta layers onto glassy carbon substrates by RF-sputtering and analyzed the electrochemical behavior in a solid state cell with Nafion<sup>®</sup> 117 as an electrolyte (Fig. 1).

This solid state cell has the advantage that the experimental conditions are close to those in a PEM-FC compared to investigations in an H<sub>2</sub>SO<sub>4</sub> electrolyte of conventional electrochemical cells. CV measurements under N<sub>2</sub> atmosphere of WC at 30°C and 60°C showed corrosion with an onset potential of about 0.5 V (DHE), attributed to an oxidation of WC to WO<sub>3</sub> and CO<sub>2</sub>. In contrast to this observation, the WC + Ta sample only showed very small anodic currents that points to a significant improved electrochemical stability at both 30°C and 60°C (Fig. 2). In voltammograms, the ORR onset potential of 0.45 V (DHE) for pure WC was shifted to 0.8 V (DHE) for the WC + Ta catalyst, pointing to an improved performance after



### Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.

**Figure 1**

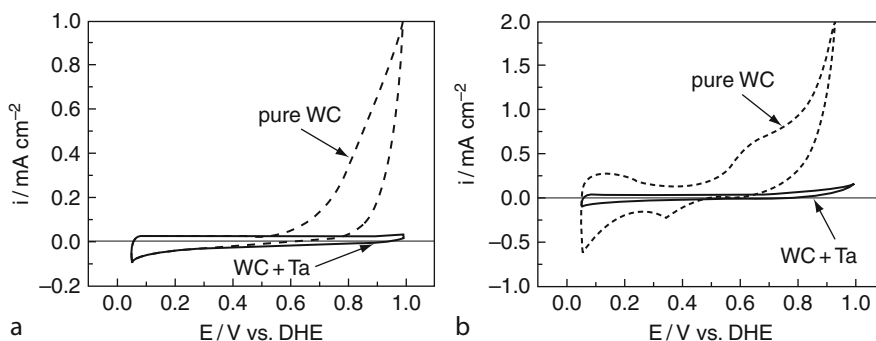
Schematic drawing of the solid state cell as used for the electrochemical measurements of WC and WC + Ta layers by Lee et al.; as counter electrode a platinum foil with platinum black was used. The figure was taken from [30], reproduced with permission of Elsevier

modification with tantalum. X-ray-induced photoelectron spectroscopy (XPS) evidenced that the WC + Ta material consisted of WC, a W-Ta alloy, TaC<sub>x</sub>, Ta and carbon (overall composition: W<sub>42</sub>Ta<sub>24</sub>C<sub>34</sub>). Among these components, the W-Ta alloy had the highest chemical stability in acidic solutions but exhibited no noteworthy ORR activity [31]. As a reason for the beneficial effect of tantalum, the authors considered that the W-Ta alloy protects the catalytically active WC-phase against corrosion that enabled the WC to reduce oxygen even at those potentials where electrooxidation was observed for the pure material.

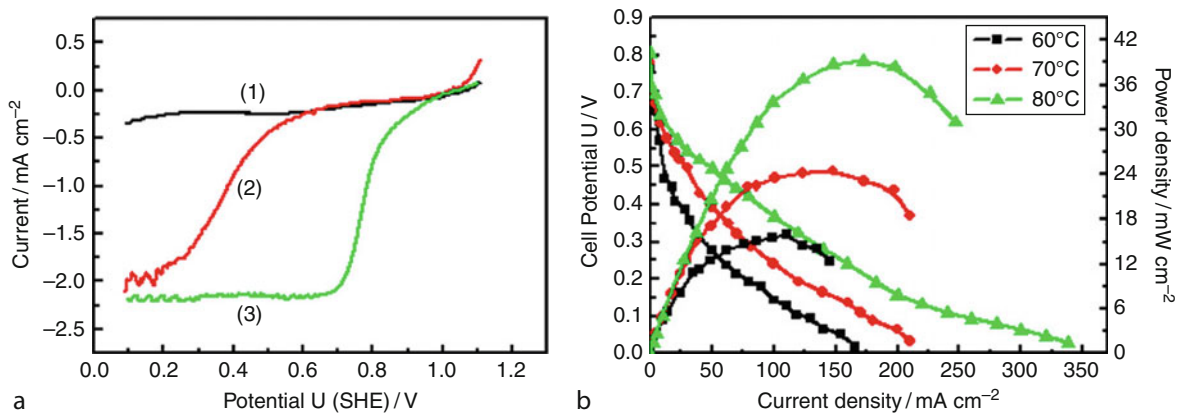
Possibly motivated by the high chemical and mechanical stability of metal nitrides as wear- and erosion-resistant coatings [32], the ORR activity of metal nitrides (Me = W, Mo) was investigated by Zhong et al. [33, 34]. The authors prepared tungsten and molybdenum nitrides as cathode catalysts by the following procedure: In the first step, a carbon black (XC-72R) impregnated with an ammonium salt of the respective metal was calcinated (500°C, N<sub>2</sub>). Subsequently, nitridation was performed by a temperature-programmed reaction under NH<sub>3</sub> gas flow (T<sub>final</sub> = 750°C and 700°C for tungsten and molybdenum). X-ray diffraction (XRD) of the final conditioned materials revealed the formation of β-W<sub>2</sub>N and γ-Mo<sub>2</sub>N nano-crystals.

From CV measurements in 0.5 M H<sub>2</sub>SO<sub>4</sub>, the authors concluded a high stability of the β-W<sub>2</sub>N/C catalyst under N<sub>2</sub> atmosphere and determined an onset potential for ORR at about 0.6 V, which is inferior compared to the reference measurement with a commercial Pt/C catalyst (onset potential of 1 V (NHE), 20 wt% Pt/C, Johnson Matthey) (Fig. 3a). In PEM-FC measurements at T = 80°C, the β-W<sub>2</sub>N/C and γ-Mo<sub>2</sub>N/C catalysts reached power densities of 39 mW/cm<sup>2</sup> (β-W<sub>2</sub>N/C, Fig. 3b) and 65 mW/cm<sup>2</sup>, respectively. Both catalysts revealed a good stability within an operating time of ≥ 60 h in galvanostatic tests (120 mA/cm<sup>2</sup> for β-W<sub>2</sub>N/C and 200 mA/cm<sup>2</sup> for γ-Mo<sub>2</sub>N/C, both at 80°C).

A slightly higher performance in PEM-FC tests was reported by Atanasoski (3 M company) using a C-N<sub>x</sub>:Fe catalyst which had been prepared by vapor deposition onto a non-woven carbon support [35, 36]. They achieved 70 mW/cm<sup>2</sup> (0.1 A/cm<sup>2</sup> at 0.7 V, 75°C, H<sub>2</sub>/air) and an open circuit potential of 0.9 V which is very



**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.** Figure 2 Cyclic voltammograms on Nafion electrolyte under  $N_2$  atmosphere, scan rate  $100 \text{ mV s}^{-1}$ , measurements were performed at  $30^\circ\text{C}$  (a) and  $60^\circ\text{C}$  (b). The figures were taken from [30], reproduced with permission of Elsevier



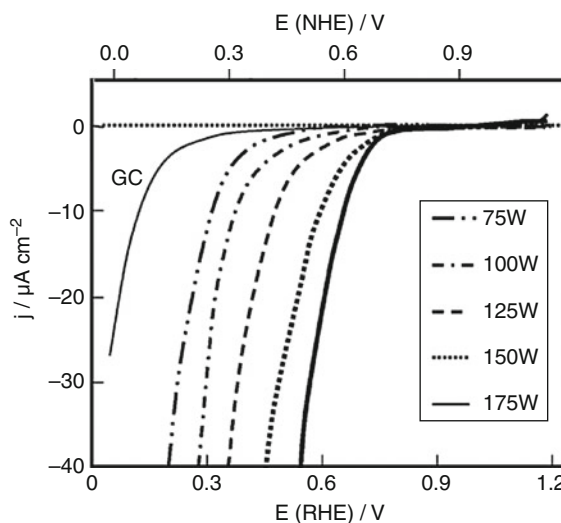
**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.** Figure 3 Cyclic voltammograms (a) for a  $W_2N/C$  ( $43 \text{ mg } W_2N$ ) in the  $0.5 \text{ M } H_2SO_4$  solution saturated with nitrogen (1) and oxygen (2), respectively. For reasons of comparison, the linear scan of a 20%  $Pt/C$  catalyst ( $12 \text{ mg } Pt$ ) in oxygen-saturated solution is given (3), all measurements were performed with a scan rate of  $5 \text{ mV s}^{-1}$  and at  $T = 25^\circ\text{C}$ . In (b), polarization curves as obtained at different temperatures using a  $W_2N/C$  catalyst (18% W) as cathode material are shown. Figures were taken from [33], reproduced with permission of Elsevier

close to that of platinum [35]. However, durability tests at  $0.65 \text{ V}$  revealed a decline of the catalyst's performance to below 10% of the initial value within 10 h. In order to improve the stability, other conductive materials like carbides, silicides, and nitrides were tested as substrates. A considerable improvement in the stability of the catalyst was found by using titanium carbide  $TiC$ . No noteworthy decrease of the fuel cell performance within 1,000 h of operation was observed. This remarkable result points out that the interaction between the support and the catalytic sites could be a very crucial factor for the optimization of catalysts.

Concerning the structure of the catalytically active centers in the  $C-N_x:Fe$  catalyst, Atanasoski et al. [36] concluded from Extended X-ray Absorption Fine Structure (EXAFS) analysis and Ultraviolet Photoelectron Spectroscopy (UPS) that the iron atoms are coordinated to nitrogen very similar to the  $FeN_x/C$  ( $x = 2, 4$ ) centers which had been detected in pyrolyzed transition metal macrocycles [37–42]. Such pyrolyzed macrocycles and related materials are currently the most active catalysts among all noble metal-free catalysts and will therefore be described in detail in section “Catalysts Prepared by Carbonization of Macrocycles” of this contribution.

Following the idea of pyrolyzed macrocycles, indeed a heat treatment of the vapor-deposited layers at 650°C led to a significantly improved ORR activity. The authors correlated this to an increased atomic order and by the introduction of a second coordination shell around the Fe atom. Similar results have been reported for iron carbon nitride layers prepared by magnetron sputtering in a combinatorial approach [43]. However, the activities are still below that of state-of-the-art pyrolyzed macrocycles. Nevertheless, these explorative works might disclose a new way to prepare similar high catalytically active structures like those known from pyrolyzed macrocycles.

Already in the 1970s, numerous transition metal oxides of perovskite- and spinel-type structure showed promising results for oxygen reduction in alkaline electrolyte; however, only moderate activities and stabilities could be achieved in acidic media. Thus, the interest to use this class of materials for the PEM-FC was nearly abandoned. In 2007, Ota's group investigated metal oxides of type  $\text{Me}_x\text{O}_y$  for  $\text{Me} = \text{Zr}, \text{Ti}, \text{Nb}, \text{Co}, \text{Sn}$ , deposited as thin layers by reactive magnetron sputtering on glassy carbon substrates [44]. Significant activities toward ORR in strong acidic media were found for zirconium oxide  $\text{ZrO}_{2-x}$  with an onset potential of about 0.64 V (NHE) followed by oxides in the sequence  $\text{Co}_3\text{O}_{4-x}$  (0.26 V) >  $\text{TiO}_{2-x} \approx \text{SnO}_{2-x}$  (0.24 V) >  $\text{Nb}_2\text{O}_{5-x}$  (0.14 V). From CV measurements ( $\text{N}_2$ -saturated 0.1 M  $\text{H}_2\text{SO}_4$  at 30°C), the authors concluded a sufficiently high electrochemical stability for all oxides in the potential range from 0.65 to 0.85 V (1.15 V for  $\text{ZrO}_{2-x}$ ), except for  $\text{Co}_3\text{O}_{4-x}$ . For  $\text{ZrO}_{2-x}$ , it was found that with increasing RF-power, the onset potential of the ORR current is shifted to more positive potentials, reaching 0.88 V (NHE) at 175 W RF-power (Fig. 4) [45]. From XPS analysis, the authors concluded that the increased amount of oxygen defects at the surface of the partially reduced  $\text{ZrO}_{2-x}$  ( $x = 0.15$ ) might serve as adsorption sites for oxygen, enhancing the ORR activity. Surface defects are discussed by many researchers to be responsible for the adsorption of oxygen molecules on metal oxide surfaces [46–48]. Additionally, the increasing number of defects could be responsible for the observed decrease in the ionization potential and the increase in the conductivity of the n-type material with increasing RF-power.



**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.**

**Figure 4**

Potential – current curves of glassy carbon (GC) and  $\text{ZrO}_{2-x}$ -layers deposited at different RF-power (as given in the figure caption) in  $\text{O}_2$ -saturated 0.1 M  $\text{H}_2\text{SO}_4$  at 30°C (scan rate: 5  $\text{mV s}^{-1}$ ). The figure was taken from [45], reproduced with permission of Elsevier

On studying  $\text{TiO}_2$  layers, an increasing ORR activity was correlated with an increasing fraction of rutile crystallites characterized by a (110) face habit leading to a decreasing ionization potential of the layers [49]. All these investigations have proven that the catalytic activity of the oxides is dependent on the crystal structure and the degree of crystallinity as well as on the electronic structure at the interface. Unfortunately, no information on the selectivity of the catalysts in the ORR was given. Mentus found on anodically oxidized Ti nanorods that in an acidic electrolyte, oxygen is predominantly reduced via a two-electron transfer pathway while in alkaline electrolytes, a four-electron transfer pathway is preferred [50]. Until now, the observed catalytic activities of these single metal oxides are smaller compared to Pt/C catalysts. Nevertheless, in contrast to Pt/C, these materials show no mixed potentials in the presence of methanol, which could make them attractive for an application in the Direct Methanol Fuel Cell (DMFC).

Based on the results with metal oxides and nitrides, research on the catalytic behavior toward ORR of metal

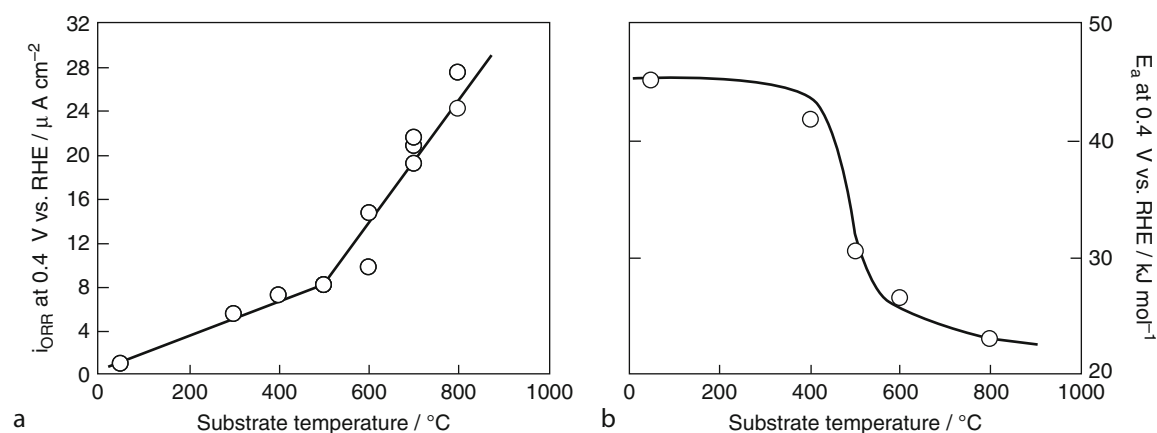
oxynitrides has been intensified in the last few years. Nitrides and oxynitrides of subgroup IV and V elements have been intensively investigated as photo-catalysts for water oxidation showing remarkable stabilities and electrochemical activities in acidic electrolytes [51, 52]. Due to their partly high metal d-band concentration at the Fermi level, these materials might also be promising as ORR catalysts.

PEM-FC tests with carbon-supported zirconium oxide  $ZrO_2$  and zirconium oxynitrides  $ZrO_xN_y$  as cathode catalysts have been reported by Liu et al. [53]. Although the maximum power density of the cell reached only one tenth of a Pt/C reference, the beneficial effect of the nitrogen incorporation into the  $ZrO_2$  lattice on the ORR activity has been clearly demonstrated. Furthermore, a good stability of the nitrogen-modified catalysts can be observed even at operation temperatures of  $80^\circ\text{C}$ . Zirconium oxynitrides have also been investigated in detail by Ota's group [54, 55]. Thin layers were prepared by RF-magnetron sputtering onto heated glassy carbon substrates. As shown in Fig. 5, they observed a strong influence of the substrate temperature on the ORR activity. Up to  $500^\circ\text{C}$ , a moderate increase of activity was observed, whereas at higher substrate temperatures, the enhancement was accelerated much stronger (Fig. 5a). This shift in activity increase was attributed to a decrease of the overall activation energy of the ORR at  $500^\circ\text{C}$  from  $42\text{ kJ mol}^{-1}$

to about  $22\text{ kJ mol}^{-1}$  (Fig. 5b); a change in the adsorption enthalpy of  $O_2$  is assumed to be responsible for the change in the activation energy [54]. In accordance with Yeager's group [56], they proposed that the rate-determining step can be attributed to a dissociative adsorption of oxygen, remaining unchanged at approximately unity over the complete investigated temperature range.

Furthermore, the ionization potential can be decreased by introducing nitrogen into monoclinic  $ZrO_2$ . However, the nitrogen concentration detected in samples annealed at  $800^\circ\text{C}$  was relatively small ( $ZrO_{1.7}N_{0.2}$ ). Similar to Liu's observations [53], CV measurements in  $0.5\text{ M H}_2\text{SO}_4$  at  $30^\circ\text{C}$  gave no hint on any electrochemical corrosion of  $ZrO_xN_y$ . For the most active  $ZrO_xN_y$  sample, deposited at a substrate temperature of  $800^\circ\text{C}$ , an onset potential for the ORR of about  $0.7\text{--}0.8\text{ V}$  was stated. More recently, an onset potential of even  $0.9\text{ V}$  had been reported for a  $ZrOCN$  catalyst which had been prepared by oxidation of  $ZrCN$  (potentials with respect to NHE) [55].

It is interesting to compare zirconium oxynitride ( $ZrO_xN_y$ ) with tantalum oxynitride ( $TaO_xN_y$ ), which was investigated by Ishihara et al. [57]. The authors reported a remarkable chemical and electrochemical stability for  $TaO_xN_y$  in  $0.1\text{ M H}_2\text{SO}_4$  at  $30^\circ\text{C}$ , that was attributed to the fact that in this oxide, tantalum is present in the highest possible oxidation state  $Ta^{5+}$  [57].



**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.** Figure 5 Relationship between the ORR current density at  $0.4\text{ V}$  (RHE) (a) and the activation energy (b), both with respect to the substrate temperature for sputtered zirconium oxynitride thin films. The figures were taken from [54] (figures 6 and 10); reproduced by permission of ECS – The Electrochemical Society

The material was prepared by nitridation of a Ta<sub>2</sub>O<sub>5</sub> powder in NH<sub>3</sub> atmosphere. The reaction product consisted of a non-stoichiometric TaO<sub>x</sub>N<sub>y</sub> (highest activity for  $x = 0.92$  and  $y = 1.05$ ) that reached a similar high onset potential as ZrO<sub>x</sub>N<sub>y</sub> of about 0.8 V (NHE) in 0.1 M H<sub>2</sub>SO<sub>4</sub>. From XRD measurements, it was concluded that the films consisted of a mixture of Ta<sub>3</sub>N<sub>5</sub> and β-TaON. The authors excluded ORR activity for Ta<sub>2</sub>O<sub>5</sub> and Ta<sub>3</sub>N<sub>5</sub> and concluded that the β-TaON which crystallizes in the same structure (monoclinic) as ZrO<sub>2</sub> is responsible for the catalytic activity. On the basis of this result, one might assume that the monoclinic ZrO<sub>2</sub> structure type has a promoting effect toward the ORR. On the other hand, in a more recent work, the authors showed good ORR activity (onset potential  $\approx 0.8$  V [RHE], type of electrolyte not mentioned) for Ta<sub>2</sub>O<sub>5</sub> which had been grown onto TaCN by a slight oxidation under low oxygen pressure of 10<sup>-5</sup> Pa at 1,273 K [58]. By using an adopted conversion electron yield x-ray absorption method (CEY-XAS), they concluded that oxygen vacancies are likely to be responsible for the ORR activity as stated by other groups (see above).

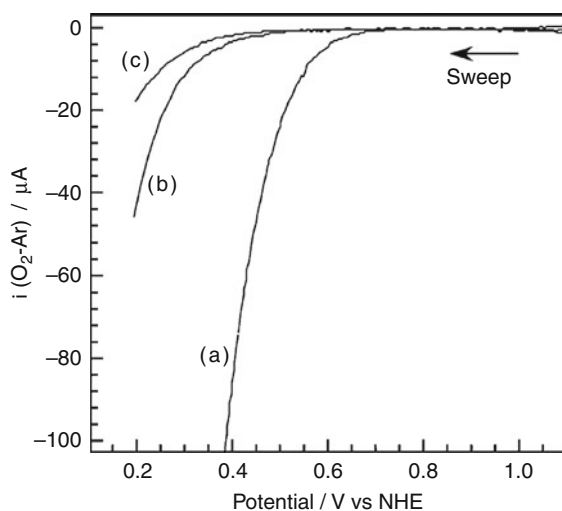
In a recent paper, Domen et al. introduced the use of an arc-plasma gun technique in a controlled O<sub>2</sub>/N<sub>2</sub> atmosphere for the preparation of niobium (oxy) nitride catalysts onto carbon blacks [59]. They showed that this technique leads to finely dispersed particles in the nanometer and sub-nanometer range which show excellent activity for the ORR compared to other non-noble metal catalysts (ORR onset potential at 0.8 V (NHE) in 0.1 M H<sub>2</sub>SO<sub>4</sub>). Also here, oxygen vacancies or reduced sites in the Nb<sub>2</sub>O<sub>5</sub> phase were proposed to be responsible for the ORR activity. Highest ORR onset potentials were observed for a N:Nb<sub>2</sub>O<sub>5</sub> catalyst (0.86 V (RHE)) which was sputtered in a reactive gas composition of N<sub>2</sub>/O<sub>2</sub> = 3:1. A further increase of the nitrogen concentrations led to the formation of less active oxynitride or nitride phases whereas lower nitrogen ratios produced less active niobium oxide (Nb<sub>2</sub>O<sub>5</sub>) phases. As no nitrogen could be detected by XPS in the most active N:Nb<sub>2</sub>O<sub>5</sub> material, only small amounts of nitrogen are obviously causing the promoting effect by doping or by forming defects in the crystalline N:Nb<sub>2</sub>O<sub>5</sub> phase. The high deposition rate of this technique enables the rapid synthesis of highly

dispersed catalysts onto high surface area substrates that is necessary for an application in fuel cells [59].

In a further paper, they reported on another alternative preparation method called polymerized complex method (PC) for niobium oxynitride supported on carbon (Nb-O-N/CB). The preparation involves the addition of carbon black (CB) during the polymerization process followed by nitridation in ammonia [60]. Such catalysts showed significantly enhanced ORR activity and also a better distribution and homogeneity of the catalyst's nanoparticles compared to a material prepared via conventional impregnation method. They attributed this effect mainly to an improved electrical contact of the catalytic active sites to the carbon substrate [60]. Obviously the use of N<sub>2</sub> instead of O<sub>2</sub> during the required heat treatment step allows the carbonization of organic precursor material which is favorable for contacting the catalyst's particles to the carbon support. Corresponding to this thesis, the authors found improved ORR current densities with increasing heat treatment temperature, probably because of an improved graphitization process.

A further doping of the Nb-O-N/CB catalyst with barium led to a considerable increase of the ORR activity by a factor of about 100 at 0.4 V (NHE) compared to the barium-free sample (Fig. 6). RRDE measurements revealed no H<sub>2</sub>O<sub>2</sub> production at potentials positive of 0.25 V (NHE), what points to a direct reduction of O<sub>2</sub> to water (four-electron pathway). Based on XRD, the authors found an increased fraction of niobium oxide Nb<sub>2</sub>O<sub>5</sub> crystallites after adding barium to the precursor. This result was confirmed by XPS where in the Ba-Nb-O-N/CB samples, an increased signal of niobium in the oxidation state Nb<sup>5+</sup> instead of Nb<sup>4+</sup> had been detected. Therefore, the authors supposed local niobium Nb<sup>5+</sup> structures close to barium atoms to be responsible for the promoting effect of the ORR activity [60].

As the examples above show, it should be pointed out that it is difficult to compare the different catalytic materials because not only the intrinsic activity but morphological parameters such as particle size, particle distribution, or electronic coupling to the substrate also play an important role in the observed activities. It has to be expected that most of the materials investigated so far can be remarkably improved in catalytic efficiency by optimizing these parameters.



### Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.

Figure 6

ORR voltammograms for (a) Ba-Nb-O-N/CB composite catalyst; (b) the Ba-free Nb-O-N/CB catalyst, and (c) bare CB. All materials were prepared by two heating steps ( $N_2$  at 773 K,  $NH_3$  at 1,123 K). Measurements in oxygen- and argon-saturated 0.1 M  $H_2SO_4$  were performed with a sweep rate of  $-5 \text{ mV s}^{-1}$ , the vertical axis ( $i_{(O_2-Ar)}$ ) denotes the difference in current under  $O_2$  and Ar atmospheres. The figure was taken from [60], reproduced by permission of ECS - The Electrochemical Society

### Non-noble Metal Catalysts with Molecular Centers

Already in 1964, Jasinski demonstrated that different metal phthalocyanines (Me: Co, Pt, Ni and Cu) can act as oxygen reduction catalysts in alkaline media [61]. During the following years, it was confirmed that several metallomacrocycles can act as ORR catalysts even in acidic environment. Later, it was found that pyrolysis leads to enormous improvements of the ORR activity even if instead of macrocycles, cheaper and less complex precursors are used [62, 63]. This will be discussed in detail in the [section "Alternative Center Generation during Heat Treatment."](#)

The current densities and stabilities of macrocycles are insufficient for fuel cell applications; nevertheless, due to their well-defined molecular structure, they are useful as model systems in order to get a better understanding of the factors that might cause the high

oxygen reduction ability of the Me-N-C catalysts prepared via pyrolysis. Therefore, this chapter will first summarize the most important factors that affect ORR activity of non-pyrolyzed macrocycles.

### Non-pyrolyzed Macrocycles

In nature, numerous redox and transport processes are associated with metallomacrocycles. As an example,  $FeN_4$ -centers of hemoglobin are responsible for the transport of oxygen in blood. In this particular case, oxygen binds reversibly to the  $FeN_4$ -unit. Besides the binding, reduction of oxygen to water requires a multi-electron transfer and the reaction with protons. Therefore, (1) all intermediate products have to be bonded strongly (but reversibly) enough to the metal center and (2) the final product of water should only have a weak bonding so that the centers are not blocked and can continue to participate in the reduction process.

The reactivity of a macrocycle will depend on the energetic position of its highest occupied molecular orbital ( $HOMO_{\text{Macrocycle}}$ ) related to the lowest unoccupied molecular orbital of the oxygen molecule ( $LUMO_{O_2}$ ). A narrow gap between both will make it reactive while a larger gap stabilizes the complex [64, 65].

Jahnke et al. investigated various macrocycles, characterized by different ligands such as nitrogen, oxygen, and sulfur ( $Me-N_4$ ,  $Me-S_4$ ,  $Me-O_4$ ,  $Me-N_2O_2$ ,  $Me-N_2S_2$ ), each under the same experimental conditions. For iron and copper, highest kinetic current densities were found if the metal center was fourfold coordinated by nitrogen ( $FeN_4$  and  $CuN_4$  centers); in the case of cobalt, the so-called Pfeiffer complex with a  $CoN_2O_2$ -center enabled the best results [62]. If oxygen or sulfur were the only ligands, the complexes were not or only poorly active toward oxygen reduction for all investigated metal centers.

Due to their low electric conductivity, macrocycles were usually impregnated on a conducting substrate, in most cases on a carbon black. Besides electron conductivity, also the active surface area is increased by the impregnation. It was found that quinone and/or carboxyl groups on the carbon surface interact electronically with the metal centers of the macrocycles. This was supposed to enhance the electronic coupling of oxygen toward the active  $MeN_4$ -site enabling higher ORR activities [66, 67]. Indeed, with Mössbauer

spectroscopy, Melendres was able to confirm that due to interaction with the carbon support, Mössbauer parameters of iron phthalocyanine (FePc) were changed confirming the different electronic structures of such interacting  $\text{FeN}_4$ -centers in comparison to  $\text{FeN}_4$ -sites in pure FePc [68].

An oxygen reduction ability was described for  $\text{MeN}_4$ -complexes with  $\text{Fe}^{2+}$ ,  $\text{Fe}^{3+}$ ,  $\text{Co}^{2+}$ ,  $\text{Ni}^{2+}$ ,  $\text{Mn}^{2+}$ ,  $\text{Ru}^{2+}$ ,  $\text{Cu}^{2+}$ , and  $\text{Zn}^{2+}$  as central ion [62, 64, 66, 67, 69–78].

Best results were found for iron, but also for cobalt, manganese, and nickel [62, 67, 79], which was explained by the 3d electron orbital occupation. It was proposed that  $\text{MeN}_4$ -centers with  $3d^6$  configuration should enable the best oxygen reduction ability [65, 77]. The related volcano plot is given in Fig. 7.

The nature and number of macrocycle substituents and axial ligands are both influencing the electron density at the metal center, leading to variations of the oxidation and/or the spin state and the electron

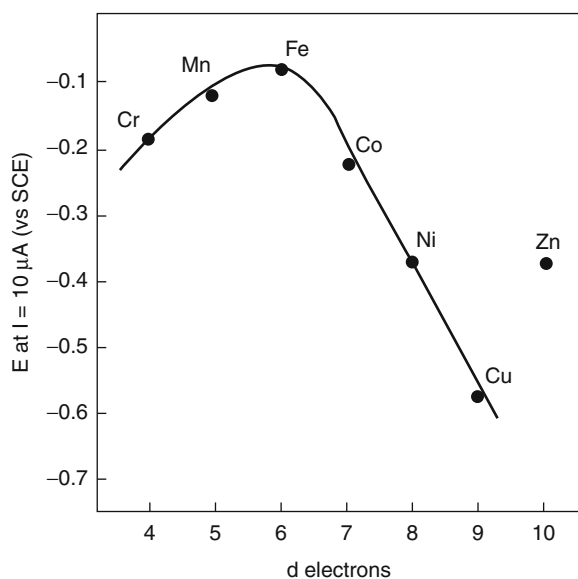
donor capacity of the complex [64, 80–84]. It was shown by STM that macrocycles with a fully occupied  $3d_z^2$  orbital are nearly ORR inactive. For the binding of oxygen, this orbital should be either empty or only partially filled [78].

Concentrating on complexes with the same metal center, it was shown that the intensity of backbonding to the complex is related to the ORR activity. As a measure for the strength of backbonding, the position of the  $\text{Co}^{3+}/\text{Co}^{2+}$  redox peak was used, i.e., depending on the macrocycle substituents, the redox peak position was shifted in the CV diagrams, the higher the redox potential the higher the kinetic current density was [85, 86].

The electron donor properties of the metal centers are affected by the type of substituents. Donating groups like phenyl or methoxy groups cause an increase of the electron density at the metal center whereas electron-withdrawing groups like sulfinyle have no or even a negative effect [62]. Induced by a higher electron density on the metal center, the oxygen molecule can be more easily activated enabling higher ORR activities [77, 87, 88].

In general, the macromolecular structure enables an extended system of conjugated  $\pi$ -electrons that provides electrons, making multi-electron transfer processes even on monomolecular centers possible. Experimentally, a direct reduction of oxygen to water was confirmed for iron and manganese  $\text{N}_4$ -macrocycles [65, 67, 79, 88–91]. For non-pyrolized cobalt complexes with single  $\text{CoN}_4$ -centers, only a two-electron reduction to hydrogen peroxide is possible. However, in 1980, it was shown that the so-called face-to-face dicobalt-porphyrins can catalyze a direct reduction of oxygen to water [92]. It was proposed that the oxygen molecule binds to two  $\text{CoN}_4$ -centers in a trans-configuration, which was suggested as bimolecular center even for pyrolyzed materials [92–97].

The formation of  $\text{H}_2\text{O}_2$  is crucial for the stability of such complexes. Hydrogen peroxide initiates a weakening or even breaking of the bonds between the tetrapyrrole core and the substituents. The resulting smaller electron donor capacity will cause a decrease of ORR activity [41, 85, 98]. On the other hand, metal ions especially in strong acidic solutions can be removed from the tetrapyrrole core by leaching. It was observed that such demetallation is a further



#### Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.

Figure 7

Volcano plot depicting the ORR onset potential of different bivalent  $\text{Me(II)}$ -TSPs for the oxygen reduction in alkaline media as a function of the number of d electrons (TSP denotes tetrasulphonatophenylporphyrin), the figure was taken from [65], reproduced with permission of Elsevier

reason for a decreased activity [71, 99–101]. As discussed in the following, both activity and stability can be enhanced by a heat treatment.

### Molecular Centers in Carbonized Materials

In 1976, it was demonstrated that the catalytic activity and stability of carbon-supported macrocycles can be enhanced significantly by a heat treatment [62, 69]. Since the 1980s, it is known that after pyrolysis, some of the metal species formed can be dissolved in acidic solution [41, 42, 63, 71, 79, 98, 102–106]. The remaining  $\text{MeN}_4$ -centers must either reduce oxygen with a higher turnover frequency or other, more active sites (e.g., N-C-sites) have to be formed. Until now, it is still under debate whether the  $\text{MeN}_4$ -centers become more active or whether the released metal acts as catalyst during the heat treatment for the formation of a certain active carbon structure (e.g., N-C-sites). It should also be noted that more than one type of catalytic center might be present in the catalysts. Possibly, both  $\text{MeN}_4$ -centers and metal-free N-C-sites could act as centers for the reduction of oxygen.

In the following, the literature will be discussed separated according to the author's assignment of active site constitution. Hence, one will also find certain metal-containing catalysts, but where the activity was assigned to N-C-sites, already discussed in the following section.

### Carbon-Based Materials with Carbon and/or Nitrogen Sites as Active Centers

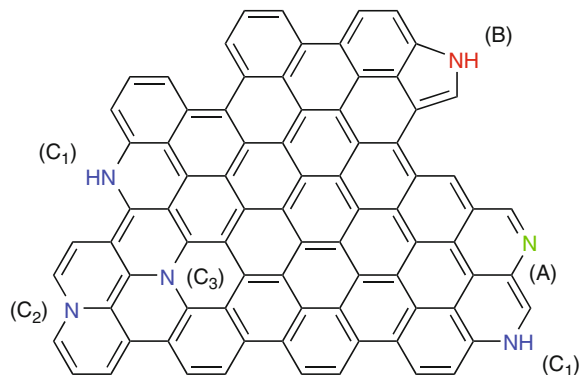
It is known that the efficiency of catalysts is not only a function of the kind of active site, but also dependent on the constitution of the carbon support, that will influence the overall electronic properties of the catalyst material [96, 107–109]. For platinum-based catalysts, it was shown that by nitrogen doping of the carbon support, the platinum particles are stronger linked to the support material compared to the not-modified carbon. The resultant smaller particles with a more homogeneous distribution enabled higher ORR activity and an improved long-term stability [107, 110].

However, beside the function as a catalyst support, it was reported that nitrogen-doped carbon itself can be active as a metal-free catalyst for the ORR. The nitrogen atoms will (1) change the electronic character of the carbon and (2) the obtained defect structure

(in comparison to the inert surface of graphene sheets) might enable an easier adsorption of oxygen and a subsequent reduction to water or hydrogen peroxide.

In 2000, Strelko et al. published theoretical calculations of the electronic character of carbon related to the integration of nitrogen, phosphorous, and boron heteroatoms [111]. In the first step, they investigated the electronic changes that depended on different positions of these atoms in the carbon matrix (i.e., pyridinic, pyrrolic, graphitic). As an example, a scheme of a graphene layer with the different type of nitrogen atoms is given in Fig. 8.

While pyridinic nitrogen atoms only contribute one electron to the  $\pi$ -electron system of the carbon, pyrrolic and graphitic nitrogen atoms increase the  $\pi$ -electron density as they contribute two electrons. Therefore, pyridinic atoms (N, P, B) should be insensitive to an improvement of the electron donor properties of the carbon. For pyrrolic and graphitic nitrogen heteroatoms, a lowering of the graphene bandgap had been calculated, enhancing the electron donor capacity in the order of  $B < N < P$  [111].



### Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.

#### Figure 8

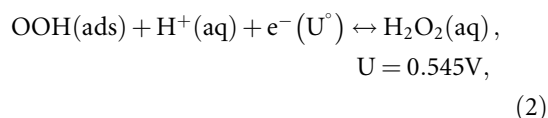
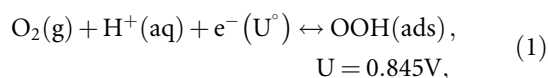
Different types of nitrogen atoms incorporated in a graphene plane. Pyridinic nitrogen atoms (A) only contribute with one electron to the  $\pi$ -electron system of the carbon matrix while pyrrolic (B) and graphitic nitrogen atoms (C) contribute by two electrons. Graphitic nitrogen atoms can be bonded in one  $\text{C}_6$ -ring ( $\text{C}_1$ ), or between two ( $\text{C}_2$ ) or three ( $\text{C}_3$ )  $\text{C}_6$ -rings, respectively. The figure was adapted from figure 3-7 of [112], Südwestdeutscher Verlag für Hochschulschriften



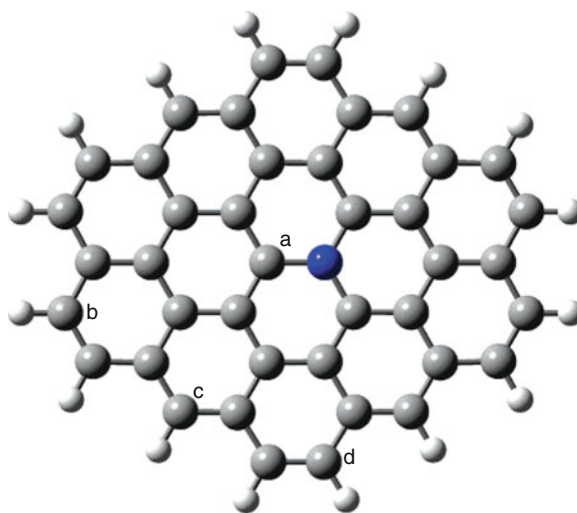
It was first stated by Wiesener that electronically modified carbon itself can act as oxygen reduction catalyst [102]. He measured ORR activity after heat treatment of  $\text{MeN}_4$ -chelates (without acid leaching). As during long-term measurements, some of the metals were dissolved without significantly affecting the ORR activity, he concluded that the metal could not be a part of active sites but catalyzes the formation of certain active N-C-sites during the heat treatment [98, 102]. Later other authors also assigned the observed ORR activity to the presence of graphitic or pyridinic N-C-sites in carbon-based materials [113–119].

The behavior of graphitic nitrogen atoms integrated in carbon was systematically investigated by Sidik et al. [120]. The catalysts were prepared starting from the soot Ketjen Black EC 300 J, which was treated in HCl and  $\text{HNO}_3$  to remove metallic impurities and then heat-treated in  $\text{NH}_3$  at  $900^\circ\text{C}$ . Induced by nitridation, ORR activity was increased but the high concentration of hydrogen peroxide that was formed (75%  $\text{H}_2\text{O}_2$  at  $U = 0.3\text{ V}$ ) indicated a predominant two-electron transfer process for this material. At higher overpotentials,  $\text{H}_2\text{O}_2$  production decreased, which was explained by an ORR process leading to water formation via two two-electron transfer processes. Cluster calculations were performed modeling the nitridation of the basal plane sheet of graphite by substitution of carbon by nitrogen atoms. For carbon atoms at different positions in the graphene sheet, the adsorption energies of reaction intermediates related to ORR were calculated as pointed out in Fig. 9.

On the basis of their calculations, the authors concluded that it is most likely that the oxygen reduction should take place on carbon atoms adjacent to in-plane N-atoms (position a in Fig. 9). With respect to the calculated potentials, for a two-step reduction



especially, the second value was in good agreement with their experimental observation of an onset potential of  $0.51\text{ V}$  (NHE). Therefore, this work evidences in experiment and theory that carbon doped with graphitic



**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.**

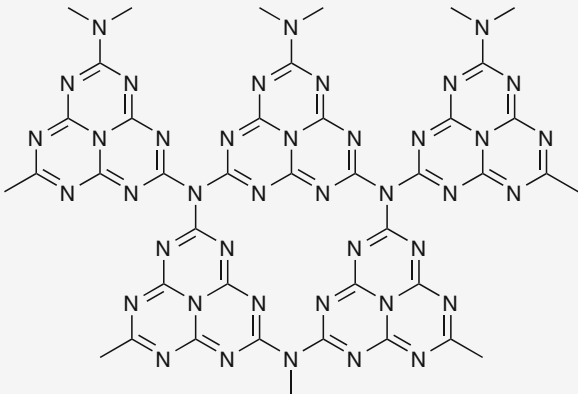
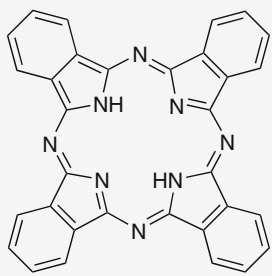
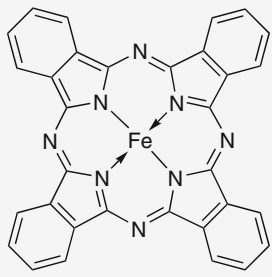
**Figure 9**

Model cluster used for the calculation of adsorption energies on (graphitic) nitrogen-doped carbon. The calculations were performed for carbon atoms at different electronic positions in the N-doped graphene plane as indicated by the letters a, b, c, and d. The figure was taken from [120], reproduced with permission of the American Chemical Society

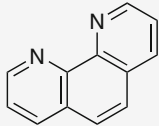
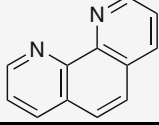
nitrogen indeed is able to reduce oxygen, but only via a two-electron pathway, to hydrogen peroxide at low onset potentials [120].

Table 1 summarizes further attempts to prepare metal-free catalysts for the ORR. Some of these materials were formed by a heat treatment of nitrogen-rich organic precursors (lines 1–2, 6); others are carbon blacks which had been heat-treated in a reactive nitrogen-rich atmosphere at appropriate temperatures (No. 4–5 in Table 1). The highest nitrogen concentration is found for carbon nitride (line 1), in which both graphitic and pyridinic nitrogen atoms have been found. Due to its good electrochemical stability, carbon nitride  $\text{C}_3\text{N}_4$  has gained interest for several catalytic applications [109, 121–123]. As this material exhibits the largest concentration of nitrogen known for N-C-based materials, it was also suggested to be of interest for the ORR. Lyth et al. investigated the electrochemical reduction of oxygen on  $\text{C}_3\text{N}_4$  which was prepared by the reaction of cyanuric chloride with sodium azide (in benzene) at  $220^\circ\text{C}$  [122]. The achieved catalytic activity was only

**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction. Table 1**  
Comparison of different carbon-based electrocatalysts and their electrochemical characteristics

	Catalyst sample		N content (wt%)	Fe content (wt%)	Onset potential V vs. NHE (@ 0.1 mA/cm <sup>2</sup> )	Lit.
1	C <sub>3</sub> N <sub>4</sub> <sup>a</sup>		60.9	–	0.4	[122, 123]
2	H <sub>2</sub> Pc/ PhR <sup>b</sup>		2.0	–	0.53	[124]
3	FePc/ PhR <sup>c</sup>		3.75	–	0.77	
4	BP 2000 NH <sub>3</sub> treated <sup>d</sup>	N:CB	1.10	–	0.4	[125]
5	Vulcan + AN <sup>e</sup>	N:CB	4.77	–	no activity	[126]

Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction. Table 1 (Continued)

	Catalyst sample		N content (wt%)	Fe content (wt%)	Onset potential V vs. NHE (@ 0.1 mA/cm <sup>2</sup> )	Lit.
6	Phen/AmOx <sup>f</sup>		n.d.	0.04	<0.0	[127]
7	Phen/AmFeOx <sup>g</sup>		n.d.	13.9	0.82	

<sup>a</sup>(cyanuric acid + sodium azide) precipitated at 220°C in benzene

<sup>b</sup>(H<sub>2</sub>Pc + phenolresin), heat-treated at 700°C for 5 h

<sup>c</sup>(FePc + phenolresin), heat-treated at 600°C for 5 h (<sup>a,b</sup>: compare Fig. 3.3)

<sup>d</sup>BP200 heat-treated in NH<sub>3</sub>/N<sub>2</sub> at 800°C for 30 min

<sup>e</sup>Vulcan pyrolyzed at 1000°C in acetonitrile-enriched Ar for 2 h, weight gain: 65%

<sup>f</sup>phenanthroline + ammonium oxalate hydrate precursor

<sup>g</sup>phenanthroline + ammonium iron oxalate trihydrate, both samples (<sup>f,g</sup>) were heat-treated at 800°C for 30 min followed by a subsequent acid leaching

slightly better compared to pure carbon black. However, after impregnation of C<sub>3</sub>N<sub>4</sub> onto the same carbon black (50/50 C<sub>3</sub>N<sub>4</sub>/CB, CB not further specified), an enhanced ORR activity was observed. The increased kinetic current density might be related to a higher active surface area and/or an improved electronic conductivity (compare Table 1, No. 1). The authors assigned the catalytic activity which was highest after a 1,000°C pyrolysis to graphitic nitrogen atoms [123]. In contrast to Wiesener [98, 102], they did not observe an improved formation of catalytically active graphitic sites by integration of iron during the heat treatment.

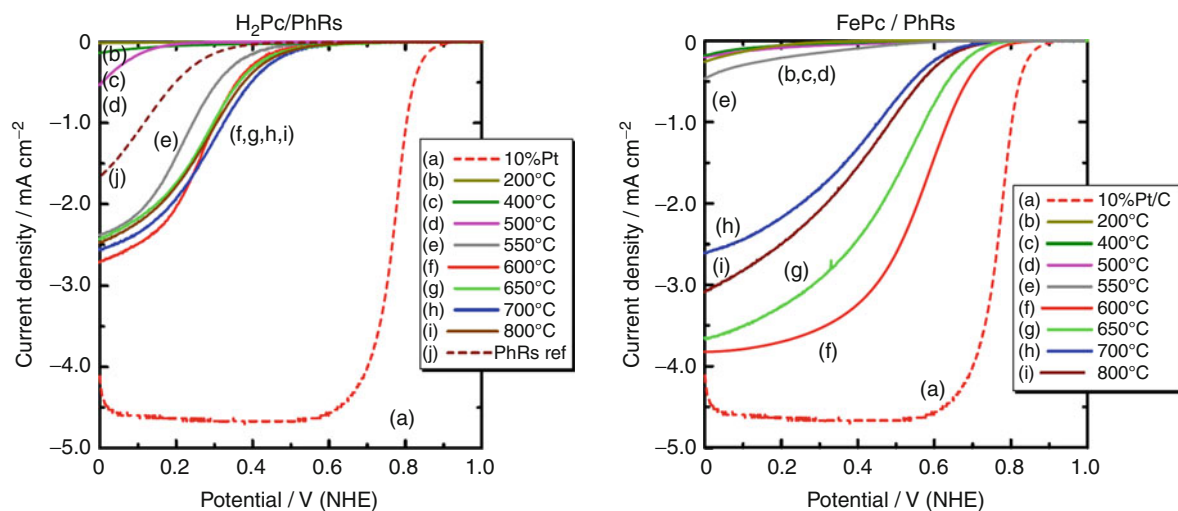
Nabae et al. [124] pyrolyzed iron phthalocyanine mixed with phenolic resin (FePc/PhR) in inert gas at different temperatures, each for five hours, and found best catalytic activity toward ORR at 600°C. A comparison to metal-free H<sub>2</sub>Pc/PhR reference samples, as shown in Fig. 10 (a: H<sub>2</sub>Pc/PhR, b: FePc/PhR), evidenced significant higher current densities and higher onset potentials for the samples prepared with iron phthalocyanine.

Similar to earlier works [104, 128], the authors observed an increased concentration of nitrogen in

the carbon structure for the iron-containing catalysts. They attributed the metal species to be more important in the enrichment of nitrogen rather than in participation in the ORR [124].

Matter et al. [129] and Maldonado and Stevenson [116] attributed the catalytic activity to pyridinic nitrogen atoms at edge-planes of “cub-stacked carbon nanotubes” [129] and carbon nanofibers [116], respectively. The authors proposed that the present metal only enhances the integration of active pyridinic nitrogen atoms.

In order to achieve a high concentration of pyridinic nitrogen atoms – without the presence of any metal source – phenanthroline was pyrolyzed in the presence of ammonium oxalate (Table 1, No. 6) [127]. The oxalate-supported pyrolysis was established for transition metal macrocycles [130–133] (compare section “Catalysts Prepared by Carbonization of Macrocycles”), but works also for metal-free phenanthroline as first shown by Herrmann et al. [134]. If ammonium oxalate is pyrolyzed together with phenanthroline, only a low active carbon-based material was obtained [127]. When the ammonium

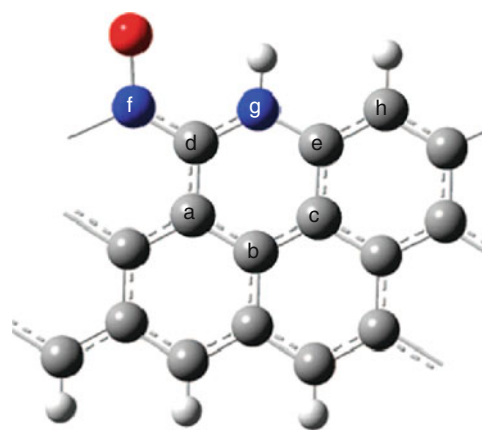


**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.** Figure 10 RDE voltammograms (0.5 M  $\text{H}_2\text{SO}_4$ , 1 mV/s, catalyst loading 0.2 mg/cm<sup>2</sup>) of (a)  $\text{H}_2\text{Pc/PhR}$  and (b) the  $\text{FePc/PhR}$  catalysts, respectively. The  $\text{FePc/PhR}$  precursor contained 3 wt% Fe, pyrolysis was performed for 5 h for all samples. All given catalysts were not acid-leached after the heat treatment. The figures were taken from [124]; reproduced with permission of Elsevier

oxalate was replaced by an ammonium iron oxalate, however, a highly active catalyst was yielded whose onset potential (at 0.1 mA/cm<sup>2</sup>) was shifted to significantly higher values compared to even the best metal-free catalyst of Table 1 ( $U = 0.8$  V compared to 0.5 V). For this iron-containing catalyst, RRDE measurements revealed a predominantly direct reduction of oxygen to water. In this case, however, the much higher activity of the Me-containing catalyst was attributed to the presence of  $\text{FeN}_4$ -centers [127].

This finding is in accord with theoretical calculations by Kurak and Anderson [135]. Using the VASP code, the authors estimated the extent to which oxygen reduction can take place on carbon, doped with pyridinic nitrogen atoms. Similar to the experimental observation, nitrogen atoms bonded near an edge of a graphene plane magnify the reactivity toward bonding radical molecules in the first step. The strong binding, however, would make large overpotentials necessary in order to enable ORR (often 1.8 V, i.e.,  $U_{\text{onset}} = -0.6$  V). Only for one specific configuration which is shown in Fig. 11, an onset potential of 0.695 V was determined, but only with hydrogen peroxide as product.

The related electron transfer mechanism for the reduction to hydrogen peroxide is shown in Fig. 12. The authors concluded that there are no evidences for a

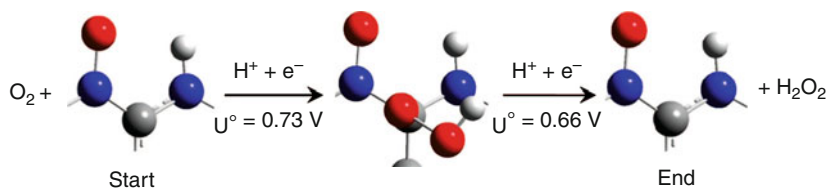


**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.** Figure 11

Proposed ON...NH edge sites on graphene sheets as catalytic sites for the oxygen reduction. The figure was taken from [135]; reproduced with permission of the American Chemical Society

direct reduction to water, but that the integration of transition metals might enhance the catalytic interaction to enable the four-electron transfer process [135].

Liu et al. prepared catalysts by a multistep preparation [115]. In the first step on either carbon or silica



**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.** Figure 12 Predicted catalytic cycle for the two-electron reduction on catalytic ON...NH edge sites in pyridinically doped carbon. The figure was taken from [135]; reproduced with permission of the American Chemical Society

supports, metal complexes were formed by the reaction of  $\text{Co}(\text{NO}_3)_2$  and  $\text{FeSO}_4$  with ethylenediamine. These precursors were heat-treated twice at  $800^\circ\text{C}$  in inert gas whereas after the first pyrolysis, an acid leaching was performed in order to remove soluble metal species. The authors came to the conclusion that in the surface region accessible for ORR, all metal-containing species should have been removed and could therefore be excluded as components of any active sites. The presence of any  $\text{FeN}_4$ -centers, stable in acidic solution, was excluded, in the first place. XPS confirmed the presence of pyridinic and graphitic nitrogen atoms. The authors proposed both to be catalytically active whereas pyridinic nitrogen atoms should have a higher activity but poorer stability (compare section “Stability”) [115].

The results presented so far have shown that in most cases, catalysts reached higher ORR activities and selectivity for a direct reduction when a metal source was present during heat treatment. This makes it difficult to finally assign the role of metals in these catalysts. Nevertheless, unquestionable is the crucial role of nitrogen atoms incorporated in the carbon structure. Even in materials, in which a metal-based  $\text{MeN}_4$ -center has undoubtedly been identified as the active site, it has been observed that graphitic nitrogen atoms (i.e., not associated to the metal) can enhance the oxygen reduction of the catalysts [136]. Further investigations are necessary to clarify the mechanism of ORR in these materials.

### Me-N-C Catalysts with $\text{MeN}_4$ and/or $\text{MeN}_{2+2}$ -Centers

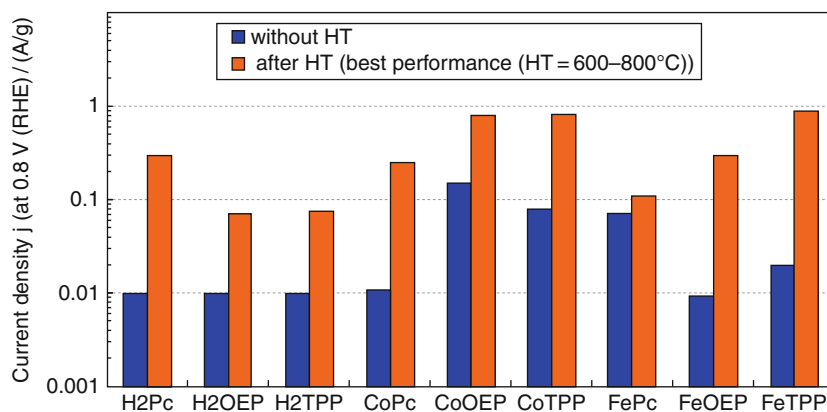
**Catalysts Prepared by Carbonization of Macrocycles**  
As already mentioned, one can enhance ORR activity and stability of macrocycle-based catalysts by performing a heat treatment [62]. This result has led

to intensive studies of the pyrolysis behavior of different  $\text{N}_4$ -chelates [37–42, 79, 100, 103, 104, 128, 137–153]. For some carbon-supported macrocycles, the effect of heat treatment on ORR activity is shown in Fig. 13.

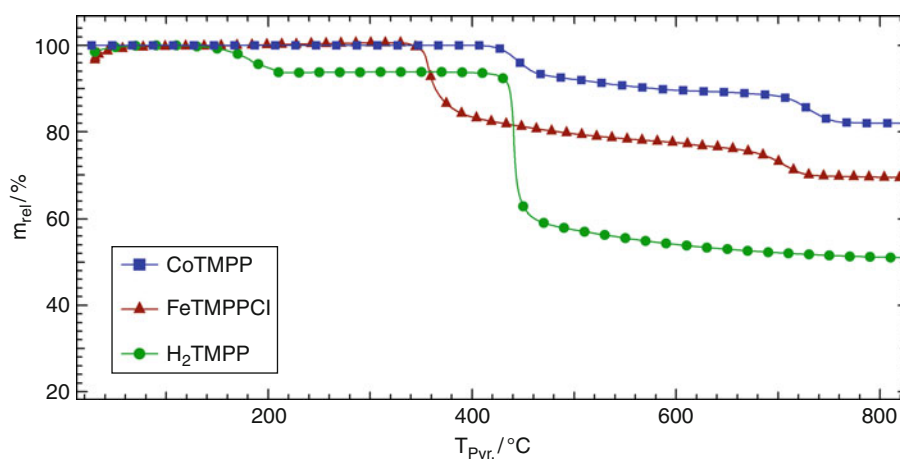
Due to the temperature-induced graphitization, an additional carbon support is not needed a priori (see below). Nevertheless, in order to increase the active surface area and thereby the current density, one can use carbon blacks as a substrate. Depending on (1) the choice of the macrocycle, (2) the pyrolysis conditions, and (3) the macrocycle loading, the maximal achievable current density will be obtained during pyrolysis in inert gas flow at temperatures between  $600^\circ\text{C}$  and  $1,000^\circ\text{C}$ . For such impregnation catalysts, an optimal loading is obtained when a double setting of the available carbon surface is given. In this approximation, the fraction of surface area which is in pores smaller than the macrocycle expansion must not be considered [145, 154]. This explains why the metal content at which highest ORR activity is found varies within the reports.

A comparison of different metal-containing and metal-free macrocycles showed that the metal centers seem to stabilize the tetrapyrrole core [104, 128, 144]. As a result, the second decomposition step is shifted to much higher temperatures compared to the metal-free porphyrin and the overall mass loss is much smaller, as it becomes apparent from Fig. 14. Furthermore, the mass fragments related to the decomposition of the tetrapyrrole core (HCN) are detected at significantly higher temperatures in the coupled mass spectrometer for metal-containing macrocycles compared to the metal-free ones [104, 112, 128, 144].

The disadvantage of the impregnation technique is that above the optimal loading, no further activity increase is possible, even a decline of the activity can be observed [130]. This fact hinders the optimization



**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.** Figure 13 Effect of heat treatment on the kinetic current density for oxygen reduction for some Co, Fe, and metal-free macrocycles (Pc: phthalocyanine, OEP: octaethylporphyrin and TPP: tetraphenylporphyrin). All samples were supported on the same carbon black; for the pyrolyzed samples, the highest kinetic current densities are given, achieved after a heat treatment at temperatures ranging from 600 to 800°C. Current values were taken from [41], figure 1



**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.** Figure 14 Thermogravimetric measurements of different tetramethoxyphenylporphyrins (TMPP). The metal-free  $H_2$ TMPP reveals a much larger mass loss starting at 450°C in comparison to both metal porphyrins. The figure was adapted from figure 5-3 of [112], Südwestdeutscher Verlag für Hochschulschriften

for FC application where higher densities of active sites are required.

For this reason, several alternative production methods were developed which enable higher site densities. In principle, the carbonization can also be reached applying low-temperature plasma. This was first shown by Herrmann et al. in 2005 [155]. The direct comparison to the pyrolyzed equivalent showed that the carbon structure of the final product contained

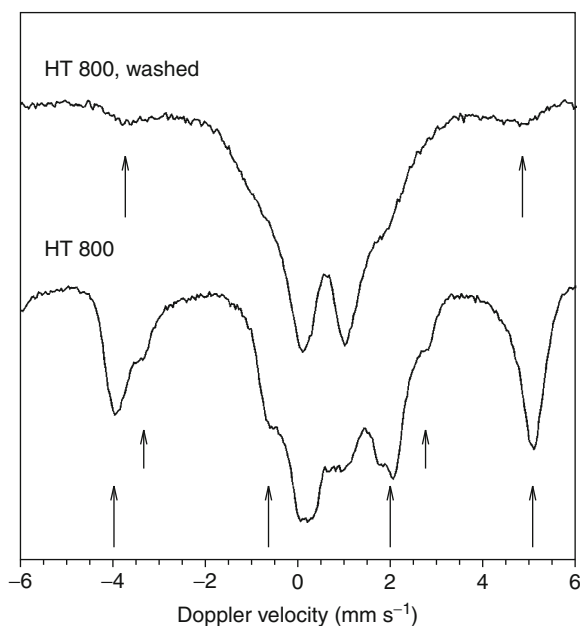
less disordered carbon phases after an Ar-plasma treatment (20 min, 250 W) compared to a pyrolysis at 700°C (for which one of the highest ORR activity was obtained) [156]. Furthermore, as only a local carbonization is reached, a sintering of the macrocycles can be prevented even without carbon support [155–157].

In order to yield higher site densities, macrocycles can either be impregnated on a template or mixed with it prior to a heat treatment. In such a case, it is always

mandatory to remove the template in a subsequent acid leaching. For instance, a sintering of macrocycle molecules is inhibited if fumed silica is used as a template [158–161]. The silica is removed by leaching in HF.

Another template-assisted method that enables high site densities is the so-called oxalate-supported pyrolysis. In this preparation approach, it is utilized that the porphyrine melts shortly before the carbonization takes place [130], whereas in the same temperature range, a decomposition of many metal oxalates occurs [132]. On the one hand, the released CO<sub>2</sub> contributes to a generation of porosity; on the other hand, the metal/metal oxide framework serves as template during the further heat treatment [39, 132, 133, 162, 163]. The advantage of this method is the achievement of a homogenous distribution of active centers over the whole catalyst material [39, 160, 162]. In a second pyrolysis, further porosities can be generated so that due to the surface increase and an increase of participating MeN<sub>4</sub>-centers, significant enhancements of kinetic current density can be achieved [39, 162]. It was shown that the addition of sulfur enables much higher current densities due to (1) an easier removal of inactive metal species during the acid leaching, (2) higher site densities, and (3) changed carbon morphology [133]. By in-situ investigation of the pyrolysis process by high-temperature X-ray diffraction (HT-XRD) and TG-MS, it was found that without sulfur addition, iron carbide formation occurs at  $T > 580^\circ\text{C}$ , causing an additional release of HCN fragments (related to MeN<sub>4</sub> decomposition). By the addition of sulfur, the formation of Fe<sub>3</sub>C is inhibited as FeS is formed instead. Therefore, a larger fraction of MeN<sub>4</sub>-centers remained intact and so higher current densities were gained. Obviously, the carbide formation (or related graphitization) should be excluded during catalyst's preparation [163].

Structural analyses showed that non-leached catalysts are often dominated by decomposition products [37, 38, 41, 42, 104, 116, 164]. In such a case, with TEM, XRD, Mössbauer spectroscopy, and EXAFS analysis, mostly metal particles, carbides, nitrides, and sometimes oxides were detected. In Fig. 15, a Mössbauer spectrum of an as-prepared catalyst (HT 800, 2 h at 800°C in Ar) and its acid-leached product (HT 800, washed) are shown.



**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.**  
**Figure 15**

Mössbauer spectra of a catalyst as prepared from carbon-supported FeTPPCL and pyrolyzed at 800°C for 2 h (HT 800) and its acid-leached product (HT 800, washed). The velocity scale is given relative to sodium nitroprusside (0.26 mm/s vs.  $\alpha$ -Fe). The figure was adapted from figure 4 of [38]; reproduced with permission of the American Chemical Society

By a sufficient leaching of the catalyst, large amounts of inorganic species can be removed without negatively affecting the kinetic current density [38, 136].

A comparison of both catalysts in Fig. 15 shows that the non-leached catalyst HT 800 is dominated by two sextets (indicated by the two different type of arrays); further Fe-containing species in the material are difficult to assign. Induced by the acid leaching the sextets nearly vanished. Thus, it becomes easier to assign the remaining species which in this case were related to different FeN<sub>4</sub>-centers [38]. Several authors have shown that even after a treatment at 1,000°C, about 30% of MeN<sub>4</sub>-centers remain intact [37, 38, 41, 42, 136]. These results show clearly that in order to understand the role of metal in the final catalysts, structural analysis should better be performed on the already acid-leached catalysts. As MeN<sub>4</sub>-centers are still

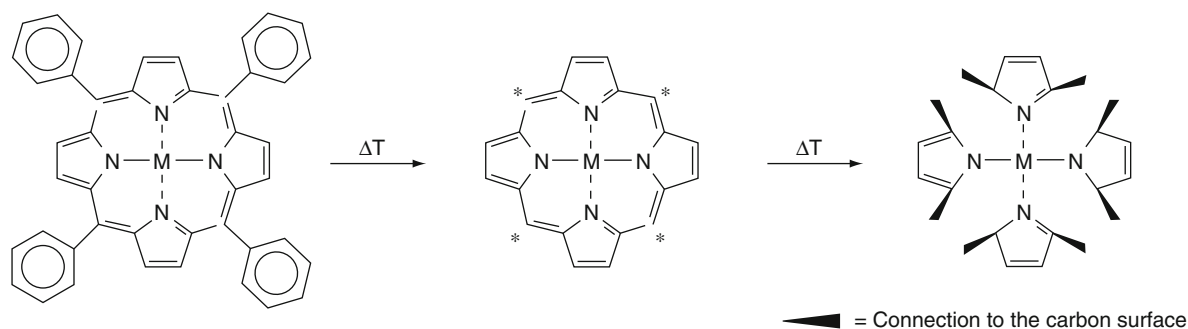
remaining, Van Veen et al. concluded that these sites are also responsible for the catalytic activity in the pyrochelates (similar to non-pyrolyzed macrocycles) [41, 42]. The authors proposed that an enhanced electronic structure is the main reason for an increased turnover frequency (TOF) of the remaining  $\text{MeN}_4$ -centers. In Fig. 16, a model for the processes occurring during heat treatment of porphyrins is shown [38]. According to this observation, even after high-temperature treatments, units of the tetrapyrrole core remain intact but interact with the carbon support.

In 2008, a direct correlation of the ORR activity and the overall concentration of all or at least one specific  $\text{MeN}_4$ -centre was proven for the first time [42]. In this work by Koslowski et al., the catalysts were produced by the oxalate-supported pyrolysis of FeTMPPCl or  $\text{H}_2\text{TMPP}$ . Variations of ORR activity and concentration of different iron modifications were obtained by different subsequent treatments whereas the heat treatment temperature was always kept constant. It was found that the obtained catalytic activity was proportional to the number of a specific  $\text{FeN}_4$ -center which is a major hint to the participation of this center in the oxygen reduction process. The correlation is shown in Fig. 17a. In a recent work, carbon-supported FeTMPPCl (FeTMPPCl/KB600) pyrolyzed at different temperatures was thoroughly structurally characterized before and after an acid leaching [136]. The results showed clearly that up to  $600^\circ\text{C}$ , the whole nitrogen remained in the system. Due to the carbonization, only a reorganization of the  $\text{FeN}_4$ -centers was observed,

while the pyrolysis at  $600^\circ\text{C}$  was sufficient to essentially improve ORR activity [39, 136].

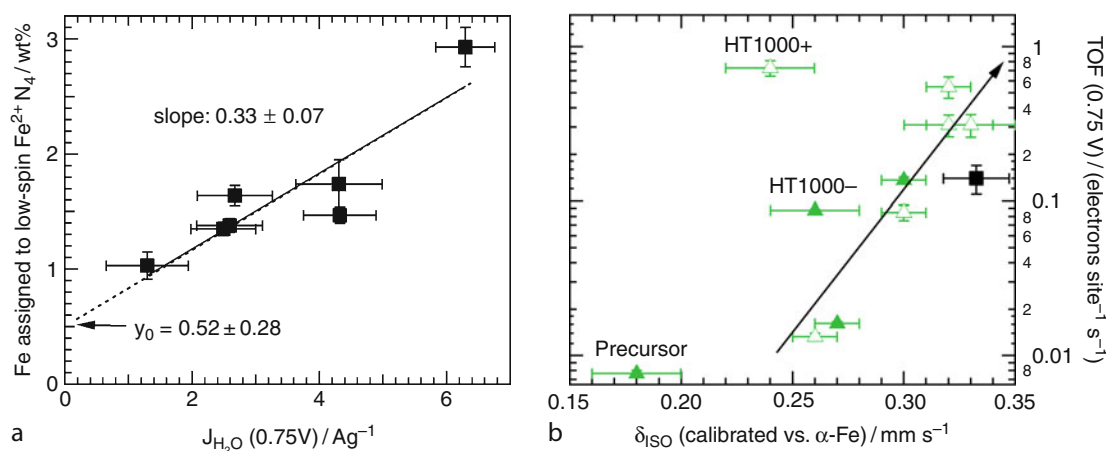
A comparison of XPS and Mößbauer spectroscopic results allows an explanation why significantly higher current densities could be achieved despite a decreasing concentration of active  $\text{FeN}_4$ -centers. Both, the N1s bonding energy and the isomer shift assigned to the active  $\text{FeN}_4$ -centers changed with the TOF [136]. The results were interpreted by an increasing bond-strength between iron and the surrounding nitrogen atoms which obviously leads to a higher 3d electron density on the metal center. As a result, oxygen might be activated faster so that higher TOFs can be obtained. A similar relation was already found for non-pyrolyzed macrocycles as discussed in section “Non-pyrolyzed Macrocycles” [65]. The relation between TOF and  $\delta_{\text{ISO}}$  assigned to the active sites is shown in Fig. 17b. According to these results, a mesomerically bonded  $\text{FeN}_4$ -center with ferrous iron in the low-spin state is assigned as catalytic active center in FeTMPPCl-based catalysts pyrolyzed in inert gas atmosphere [39, 136]. Apart from these findings, graphitic nitrogen atoms seem to enhance the electron donor capacity, thus enabling higher ORR activities [136].

A related positive effect emerged also after a second pyrolysis of Me-N-C catalysts in ammonia [112, 160, 162, 165–167]. Since a huge increase in the concentration of pyridinic nitrogen atoms was observed, it was suggested that either different, but more active sites were formed or an enhanced electronic structure of the carbon matrix enabled the higher ORR activity



**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.** Figure 16 Visualization of the carbonization process of a porphyrin during the heat treatment. The figure was taken from [38]; reproduced with permission of the American Chemical Society. During the first decomposition step of porphyrins (compare figure 14), the substituents are released. In the following heating process, the active centers are integrated into or onto the carbon framework





**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.** Figure 17 Correlation of the concentration of low-spin Fe<sup>II</sup>N<sub>4</sub>-centers with the ORR current density for different Fe-N-C catalysts (a). Relation between turnover frequency (TOF) and isomer shift δ<sub>ISO</sub>. (b) The isomer shift δ<sub>ISO</sub> is related to the electron density at the iron nucleus. In this work, a higher δ<sub>ISO</sub> is in accord with an increased 3d electron density. The precursor and impregnation catalysts prepared at 1,000°C (HT1000+: acid-leached, HT1,000-: without acid leaching) are indicated. For reasons of comparison, the point ■ related to the catalysts discussed in [39] is added (average values of δ<sub>ISO</sub> and TOF). Data were taken from [39] and [136], reproduced with permission of the American Chemical Society (for figure a), and permission by ECS – The Electrochemical Society (for figure b), respectively

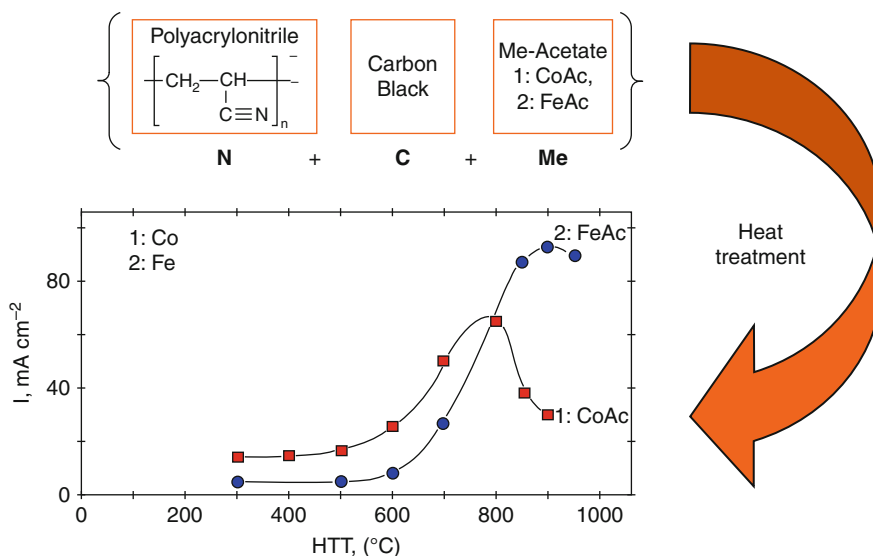
[115, 117, 118, 126, 129, 165, 168–173]. In general, the utilization of NH<sub>3</sub> allows an alternative generation of catalysts where instead of complex macrocycles, simple and cheap substances can be used. This matter will be discussed in the following.

**Alternative Center Generation During Heat Treatment** In 1989, Gupta and coauthors showed that highly active catalysts can be prepared from less complex molecules [63]. In their work, they impregnated a carbon black with polyacrylonitril (PAN) and an iron or cobalt acetate. The precursors were heat-treated at different temperatures and the ORR activity was measured. It was found that one can generalize the preparation of Me-N-C catalysts: Whenever a metal precursor is heat-treated with nitrogen and carbon sources at temperatures of ≥600°C (Co) or ≥700°C (Fe), an active catalyst can be obtained. A scheme of their preparation route and the achieved ORR activities (as a function of pyrolysis temperature) are given in Fig. 18.

Since then, this approach was adapted by several working groups [43, 97, 113, 115–118, 126, 152, 165, 166, 168, 171, 173–186]. Extensive studies were made

by Dodelet's group, who tested a wide gamut of metal, nitrogen, and carbon sources, in order to find highest current densities by this screening technique. It could be demonstrated that metal acetates and ammonia are especially suited as precursors to achieve highly active catalysts [179, 187]. Besides ammonia, other nitrogen precursors such as acetonitrile (AN), metal-free N<sub>4</sub>-macrocycles, phenanthroline (Phen), polypyrrole (PPy), polyanniline (PANI), and others can equally be used as demonstrated by several groups [113, 126, 127, 134, 165, 174, 176–178, 186, 188, 189].

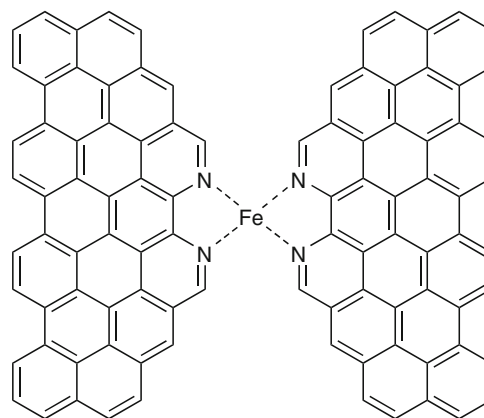
Varying the metal acetate in the precursor, the obtained ORR activity was studied by He et al. [187]. It was found that the kinetic current density increases in the order Ni (3d<sup>8</sup> 4s) ≈ Cu (3d<sup>10</sup> 4s) < Mn (3d<sup>5</sup> 4s<sup>2</sup>) < Cr (3d<sup>5</sup> 4s) << Co (3d<sup>7</sup> 4s<sup>2</sup>) < Fe (3d<sup>6</sup> 4s<sup>2</sup>). If the ORR activity is plotted versus the 3d orbital occupation a volcano plot with a maximum at 3d<sup>6</sup> will be obtained (For the ease of comparison, the electron configuration of the metals is given in brackets). It should be pointed out that this order nearly reflects the relation between ORR onset potentials and the number of 3d electrons as shown for non-pyrolyzed porphyrins in Fig. 7.



**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction. Figure 18** Scheme of Gupta's precursor selection (*top*) and ORR activity of catalysts as a function of the heat treatment temperature and metal precursor, 1: CoAc, ■ and 2: FeAc, ● (*bottom*). ORR activity was measured in 4 M NaOH. The temperature-dependent current curve was taken from [63], Figure 3; reproduced with permission of Springer Publishing Group

It has been observed that besides the type of metal and its content, the fractions of micropores and pyridinic nitrogen atoms play a crucial role for the active site formation. Both are generated during the ammonia treatment [168, 171, 180, 182, 187]. The authors assumed that  $\text{MeN}_{2+2}$ -centers were built in micropores, a scheme of the center is given in Fig. 19 [168]. Experiments performed with the goal to increase the amount of active centers by annealing the non-porous carbon black impregnated with iron acetate in the presence of  $\text{NH}_3$  failed. Even catalysts with more than 0.2 wt% Fe turned out to reveal only smaller turnover frequency. Above 2 wt% iron, the activity dropped remarkably [172].

Mössbauer analysis of such catalysts evidenced that besides less active  $\text{FeN}_4$ -centers (similar to those in heat-treated porphyrin-based catalysts), a new type of  $\text{FeN}_4$ -centers was found [190]. This high-spin  $\text{Fe}^{2+}$ -center was not present in any heat-treated  $\text{N}_4$ -macrocycle-based catalyst, making it most probable that it is related to the proposed active site. In conclusion, the  $\text{FeN}_{2+2}$ -centers differ from porphyrin-based ones by the different spin state, the pyridinic coordination of the iron ion and the placement in micropores. The different electronic state might explain why the



**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction. Figure 19**

$\text{FeN}_{2+2}$ -center embedded in a micropore, as assumed by Charretour [168]. The figure was taken from [112], figure 3-9; Südwestdeutscher Verlag für Hochschulschriften

$\text{FeN}_{2+2}$ -centers can reach much higher turnover frequencies in their "highly active state" (please note: the electron density on iron is higher for  $\text{Fe}^{2+}$ , HS compared to  $\text{Fe}^{2+}$ , LS); on the other hand, these centers can

be deactivated, e.g., in contact with acid, as will be discussed in the section “Stability” [190, 191].

Very high activities are obtained when phenanthroline is used as nitrogen source [134, 165, 178]. In 2009, such catalysts exhibited the highest catalytic activity within the system Fe-N-C. The authors assigned their preparation method as pore-filling method (PFM) because especially this step within the preparation process seems to play a crucial role to achieve a high ORR activity [165]. In order to obtain the precursor, a microporous carbon was impregnated with phenanthroline and iron acetate. In the next step, a low-energy ball milling was performed with the aim to press iron acetate and phenanthroline into the micropores. The precursor was pyrolyzed two times (first 1,050°C in Ar, second at 950°C in NH<sub>3</sub>) [165]. Highest ORR activities were achieved when the total weight loss from both pyrolysis steps amounted to the weight percentage of phenanthroline, the best having a phenanthroline content of 50 wt% [165, 192].

A homogeneous distribution of active sites was obtained when the phenanthroline is pyrolyzed in the presence of iron oxalate. Such catalysts exhibited an excellent performance in RDE measurements, even without sulfur addition [134]. As indicated by a change of color, already during the mixing of the precursors, a complexation of iron with phenanthroline occurs. These catalysts show similar structural features in Mössbauer spectroscopy and exhibit similar iron contents as those catalysts prepared by the oxalate-supported pyrolysis of porphyrins.

A further preparation approach uses nitrogen-bearing monomers to synthesize catalytic active sites [177, 186, 189], e.g., when pyrrole is impregnated on a carbon black in the presence of a metal source. An oxidation agent induces the polymerization process. Without any heat treatment, only small ORR activity was obtained; however, this catalyst showed a stable performance of more than 500 h in potentiostatic fuel cell tests [177]. Although a heat treatment enhanced the catalyst's activity (similar to the pyrolysis of macrocycles), the stability was moderate. Nevertheless, recent results have shown that by further optimization of the preparation method (including a heat treatment), both, a high activity and a remarkable stability, could be realized [189]. This fact will further be

discussed in section “NNMC and Their Potential for PEM-FC Application.”

The preparation of catalysts by utilization of monomers with a subsequent polymerization reaction and heat treatment was also applied by other groups [159, 160, 193, 194]. Dahn's team impregnated silica with pyrrole and iron chloride. The polymerization was induced in a hot acid steam. Directly after complete polymerization, the sample was pyrolyzed in Ar at 900°C. The silica template was removed afterward by a doubled leaching in 5% and 40% HF, respectively [159, 160].

Relatively high iron contents are found for self-supported iron-polypyrrole catalysts which were prepared by spray pyrolysis. The authors proposed high site densities; however, comparatively low current densities with respect to other Fe-N-C catalysts make it most probable that only a small fraction of the overall iron is bonded in active sites. Nevertheless, the shape of these catalysts is interesting because porous carbon spheres with diameters of 100 up to 1,000 nm are formed [194].

In conclusion, it can be summarized that active sites in Me-N-C catalysts can be prepared via several different preparation routes. Optimization of each parameter (e.g., amounts of precursors [metal, nitrogen, and carbon], ratios of precursors, annealing temperature, and post treatment) is necessary in order to achieve highest ORR activity. If a heat treatment is performed, one will always achieve a mixture of different metal modifications, whereas it seems that only a fraction of the modifications formed is connected with ORR activity. This observation supports the importance of a subsequent acid leaching and of highly sensitive analysis techniques that can distinguish between the different phases, especially between FeN<sub>4</sub>-centers of different oxidation and spin states.

## NNMC and Their Potential for PEM-FC Application

With respect to the alternative materials discussed so far, Fe-N-C catalysts seem to be the most promising cathode catalysts for FC application [165, 195]. For that reason, the discussion is focussed on this material system in this chapter. Alternative fuel cell catalysts have to be (1) simple in their preparation, (2) economical in production, and (3) they should generate an

adequately high volumetric current density ( $J_{\text{vol}}$  in  $\text{A}/\text{cm}^3$ ). In order to meet the DOE-target value of the volumetric current density, a certain site density ( $S_{\text{D}}$  in  $\text{sites}/\text{cm}^3$ ) and a sufficiently high turnover frequency (TOF in  $\text{electrons sites}^{-1}\text{s}^{-1}$ ) have to be gained to enable commercial application. Besides, a catalyst stability of several 1,000 h is expected, even under high load conditions of the fuel cell.

Today, catalysts that reach high volumetric current densities often exhibit only low stability while different, less active materials perform nearly stable, even over periods of weeks. Therefore, activity and stability will be discussed separately. The section “Future Direction” will give a final outlook regarding the commercial applicability of NNMC.

### Activity

Up-to-date non-noble metal catalysts only partially fulfill the demands in activity for a fuel cell application. To boost activity, the catalyst loading on the cathode can be increased. It was noticed, however, that above a certain layer thickness, mass transport properties hinder a further increase of the current density. For this reason, the specific volumetric activity has to be improved.

The target values for 2010 (the year in which this article was written) and for 2015 are 10% ( $130 \text{ A}/\text{cm}^3$ ) and 25% ( $325 \text{ A}/\text{cm}^3$ ), respectively, of the volumetric ORR activity of a commercial Pt/C catalyst [160, 166, 195]. Standard conditions are fulfilled if the measurements are performed at  $80^\circ\text{C}$  ( $T_{\text{CELL}} = 80^\circ\text{C}$ ) with oxygen and hydrogen partial pressures of 1 bar ( $p_{\text{O}_2} = 1 \text{ bar}$ ,  $p_{\text{H}_2} = 1 \text{ bar}$ ) and a relative humidity of  $\text{RH} = 100$ .

In Table 2, different catalysts are compared with respect to their volumetric ORR activity, site density, kinetic current density, and turnover frequency. The knowledge of these different parameters enables the identification of the factors that have to be further optimized. The catalysts are sorted according to their achieved volumetric current density calculated from fuel cell measurements (column I, Eq. 3). The so far best Fe-N-C catalyst is listed below the DOE-target value followed by the other materials given in the sequence of decreasing volumetric activity. If the authors did not provide any magnitude of the volumetric

current density, the value was calculated under the assumption of a catalyst density of  $\rho = 0.4 \text{ g}/\text{cm}^3$  [172, 196].

$$\text{Volumetric current density } J_{\text{vol}} \text{ (in } \text{A cm}^{-3}\text{):}$$

$$J_{\text{vol}} = \frac{J_{\text{A/g}}}{\rho_{\text{g}/\text{cm}^3}}, \quad (3)$$

Presently, none of the Fe-N-C catalysts meet the target value of  $J_{\text{vol}}$ , even if some of the given data were extrapolated to 0.8 V from the Tafel region of the polarization curve. For highly active catalysts even at 0.8 V, some mass transport limitation is visible so that the measured values of the current density are smaller compared to the extrapolated ones.

An increase of the volumetric current density can be obtained either by enhancing the site density  $S_{\text{D}}$  and/or the turnover frequency TOF. Column C in Table 2 lists the site densities of the catalysts as far as the metal content is known. The calculations were made under the assumption that each metal atom is coordinated in an active  $\text{MeN}_4$ - or  $\text{MeN}_{2+2}$ -center.

$$\text{Site density } S_{\text{D}} \text{ (in sites}/\text{g}_{\text{cat}}^{-1}\text{): } S_{\text{D}} = \frac{[\text{Me}]_{\text{wt}\%}}{100 \cdot M(\text{Me})_{\text{g}/\text{mol}}} N_{\text{A}}, \quad (4)$$

In the equation,  $[\text{Me}]_{\text{wt}\%}$  is the concentration of metal species in the catalyst,  $M(\text{Me})_{\text{g}/\text{mol}}$  is the molar mass (if several metal species are present, the average related to the contents of different metals is used), and  $N_{\text{A}}$  represents Avogadro's number. Since the density for many of the catalysts is unknown, a weighting related to the volume was neglected.

As described in the section “Non-noble Metal Catalysts with Molecular Centers,” it is still controversially discussed, what structural unit (or units) in Me-N-C catalysts causes the oxygen reduction reaction. Furthermore, it was shown for different catalysts that besides the presumably active  $\text{MeN}_4$ -center also inactive metal-containing modifications exist [39, 40, 117, 129, 131–133, 136, 162, 163, 170, 172, 175, 185, 197, 198]. Consequently, the given site density  $S_{\text{D}}$  can merely be regarded as a rough approximate. Nevertheless, it helps to estimate if the site density is a distinct value which has to be adapted.

From the site density  $S_{\text{D}}$  and the kinetic current density  $J_{\text{K}}$  (at 0.8 V), one can determine the turnover frequency TOF (0.8 V), which is given in Table 2 for

**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction. Table 2** Comparison of several catalysts concerning their site density  $S_D$  (column C), their kinetic current densities (columns D and E), turnover frequency (columns G and H), and their achieved Volumetric current density (column I). The values in Columns D – I are all given for a potential of 0.8 V. To enable a comparison, for several catalysts values were calculated from the original data as described in the beginning of the chapter Activity

	A	B	C	D	E	F	G	H	I	J
	Shortname	Lit.	$S_D$	RDE	FC	FC/RDE	RDE	FC	FC	Description
			$\times 10^{20}$ /sites/g <sub>cat</sub>	J (0.8 V) /A/g			TOF (0.8 V)/ electrons/ (site s)	vol curr. J <sub>v</sub> /(A/cm <sup>3</sup> )		
1	Target NNMC 2010	DOE							130	
2	Target NNMC 2015	DOE							325	
3	CM-Fe-C	208		–		–	–		127	CM: Cyanamide, Fe-source: FeSO <sub>4</sub> *7H <sub>2</sub> O, C-source: sucrose, leaching in KOH
4	PFM1: Fe/Phen/BP-HT1050Ar-HT950NH <sub>3</sub>	192	2.2	–	429	–	–	12.42	99	optimized PFM-catalyst, Fe-source: iron acetate
5	PFM2: Fe/Phen/BP-HT1050Ar-HT950NH <sub>3</sub>	165	2.2	–	246	–	–	7.12	98.3	catalyst after optimization, Fe-source: FeAc
6	PANI-Fe/EDA-Co-C	208	–	13.75	90	6.55	–		72	PANI-Fe/EDA-Co-C
7	PFM4: Fe/PTCDA/BP-HT1050NH <sub>3</sub> ("M786")	160	0.8	3	80	26.67	0.23	6.2	32	HT 1050°C, 5 min NH <sub>3</sub> , Fe-source: FeAc
8	Fe/Fe/S-HT800N <sub>2</sub> -HCl-BM-HT800NH <sub>3</sub>	209	5.1	0.71	76.4	107.61	0.01	0.93	30.6	Fe/Fe/S denotes mixture of FeTMPPCI + iron oxalate + sulfur, Leaching in 1 M HCl, 15 min ball milling (BM)
9	Fe/Fe/S-HT800N <sub>2</sub> -HCl-HT800NH <sub>3</sub>	209	4.3	4.48	31	6.92	0.08	0.45	12.4	Fe/Fe/S denotes mixture of FeTMPPCI + iron oxalate + sulfur, Leaching in 1 M HCl
10	3M-NiANI-HT_NH <sub>3</sub> *-HT_NH <sub>3</sub> *	166	1.3	–	30	–	–	1.45	12	iron-source: FeCl <sub>3</sub> , NiANI: Nitroanilin, NH <sub>3</sub> *: gas mixture of 25% NH <sub>3</sub> and 75% N <sub>2</sub> , HT at 800–1,000°C, same T for both steps

Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction. Table 2 (Continued)

	A	B	C	D	E	F	G	H	I	J
	Shortname	Lit.	$S_D$	RDE	FC	FC/RDE	RDE	FC	FC	Description
			$\times 10^{20}$ /sites/ $g_{cat}$	J (0.8 V) /A/g			TOF (0.8 V)/ electrons/ (site s)		vol curr. $J_v$ /(A/cm <sup>3</sup> )	
11	Fe/Fe/S-HT800N <sub>2</sub> - HCl-HT800N <sub>2</sub> /H <sub>2</sub> - HCl-HT800CO <sub>2</sub> ("UK63")	160	3.4	18	20.7	1.15	0.33	0.38	8.3	Leaching in 1 M HCl; N <sub>2</sub> /H <sub>2</sub> : Forming Gas 10% H <sub>2</sub> ; Fe/Fe/S denotes mixture of FeTMPPCl + iron oxalate + sulfur
12	PFM3: Fe/Phen/BP- HT1050Ar	165	1.0–2.0	–	13.8	–	–	0.4–0.8	5.5	only heat treatment in Ar, Fe-source: FeAc
13	Co/Fe/S-HT800N <sub>2</sub> - HCl-HT800N <sub>2</sub> /H <sub>2</sub> - HCl-HT800CO <sub>2</sub> ("UK65")	160	3.3	10	10.3	1.03	0.19	0.19	4.1	For leaching 1 M HCl; N <sub>2</sub> /H <sub>2</sub> : Forming gas with 10% H <sub>2</sub> ; Co/Fe/ S denotes mixture of CoTMPP + iron oxalate + sulfur
14	(PANI-Fe <sub>3</sub> Co- KB300)-HT900N <sub>2</sub> - H <sub>2</sub> SO <sub>4</sub>	186	2.7	1.3	9.8	7.54	0.03	0.23	3.9	Fe- and Co-sources: hydrated Fe/Co- sulfates, leaching in 0.5 M H <sub>2</sub> SO <sub>4</sub>
15	$\mu$ PM1: Fe/npCB- HT950NH <sub>3</sub> ("FC280")	160	0.3	4.2	5.8	1.38	0.92	1.27	2.3	$\mu$ PM: Micropore method; Fe-source: FeAc; npCB: non- porous carbon black
16	GAdFeCu- HT1000Ar-H <sub>2</sub> SO <sub>4</sub> ("GAdFeCu")	160	3.0	0.55	5.7	10.36	0.01	0.12	2.3	G: Glucose, Ad: Adenin, Fe- and Cu-source: hydrated Fe/Cu-gluconate
17	"DAL900c": (DAL900a + Fe)- HT900NH <sub>3</sub>	160	1.5	1.1	5	4.55	0.05	0.21	2	DAL900a + Fe: impregnation of DAL900a with FeAc previous to HT
18	CoTMPP/SiO <sub>2</sub> - HT700N <sub>2</sub> -KOH ("CoTMPP700")	160	6.8	1	3.2	3.20	0.01	0.03	1.3	CoTMPP was impregnated on porous silica previous to HT
19	"DAL900a": (FeCl <sub>3</sub> / SiO <sub>2</sub> + PPy/HCl)- HT900Ar-HF	160	0.5	0.4	2.7	6.75	0.05	0.36	1.1	PPy/HCl: Polymerization of pyrrole by HCl- vapor; HF: leaching in 5% and 40% HF- solution

Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction. Table 2 (Continued)

	A	B	C	D	E	F	G	H	I	J
	Shortname	Lit.	$S_D$	RDE	FC	FC/RDE	RDE	FC	FC	Description
			$\times 10^{20}$ /sites/g <sub>cat</sub>	J (0.8 V) /A/g		TOF (0.8 V)/ electrons/ (site s)	vol curr. J <sub>v</sub> /(A/cm <sup>3</sup> )			
20	Fe/Fe/S-HT800N2-HCl-BM	209	4.3	0.4	2.3	5.75	0.01	0.03	0.9	Fe/Fe/S denotes mixture of FeTMPPCl + iron oxalate + sulfur; BM: 15 min ball milling
21	Hem-HT600Ar-G-HT900CO <sub>2</sub> *-H <sub>2</sub> SO <sub>4</sub> ("CHb200900")	160	1.2	0.4	2.3	5.75	0.02	0.12	0.9	Hem: grounded Hemoglobin; G: grounded, CO <sub>2</sub> *: gas mixture of 10% CO <sub>2</sub> and 90% Ar, leaching in 0.5 M H <sub>2</sub> SO <sub>4</sub>
22	Co-PPy-C (Vulcan + PPy/H <sub>2</sub> O <sub>2</sub> ) + Co (NO <sub>3</sub> ) <sub>2</sub> /(NaBH <sub>4</sub> + NaOH)	177	4.1	-	1.8	-	-	0.03	0.72	Vulcan was impregnated with pyrrole, polymerization was done by H <sub>2</sub> O <sub>2</sub> ; product was mixed with Co(NO <sub>3</sub> ) <sub>2</sub> , which was reduced by NaBH <sub>4</sub> + NaOH
23	Fe/Fe/S-HT800N <sub>2</sub> -HCl	209	3.9	0.19	1.7	8.95	<0.01	0.03	0.7	For leaching 1 M HCl; Fe/Fe/S denotes mixture of FeTMPPCl + iron oxalate + sulfur
24	3 M-NiANI-HT_NH <sub>3</sub> *	166	1.3	-	1	-	-	0.05	0.4	iron-source: FeCl <sub>3</sub> , NiANI:Nitroanilin, NH <sub>3</sub> *: gas mixture of 25% NH <sub>3</sub> and 75% N <sub>2</sub> , HT at 800-1,000°C
25	FeCl <sub>3</sub> /SiO <sub>2</sub> + PPy/HCl-HT900Ar-HF	159	2.2	1.3	1	0.77	0.04	0.03	0.4	PPy/HCl: Polymerization of pyrrole by HCl-vapor; HF: leaching in 5% and 40% HF-solution
26	Fe-PPy-MS	194	2.9	0.14	0.4	2.86	0.02	<0.01	0.16	Polymerization of pyrrole + K <sub>3</sub> [Fe(CN) <sub>6</sub> ] in H <sub>2</sub> O + H <sub>2</sub> O <sub>2</sub> induced by FeCl <sub>3</sub> ; US spray pyrolysis (800°C, 2.4 MHz), leaching in 10% HF-solution

Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction. Table 2 (Continued)

	A	B	C	D	E	F	G	H	I	J
	Shortname	Lit.	$S_D$	RDE	FC	FC/RDE	RDE	FC	FC	Description
			$\times 10^{20}$ /sites/ $g_{cat}$	J (0.8 V) /A/g			TOF (0.8 V)/ electrons/ (site s)		vol curr. $J_v$ /(A/cm <sup>3</sup> )	
27	((CB-HCl-HNO <sub>3</sub> ) + Co/Fe/N)-HT900Ar-H <sub>2</sub> SO <sub>4</sub>	119	6.5	0.28	0.1	0.36	<0.01	<0.01	<0.01	CB-HCl-HNO <sub>3</sub> : carbon was first leached in conc. HCl then in conc. HNO <sub>3</sub> ; Co/Fe/N: Co(NO <sub>3</sub> ) <sub>2</sub> + FeSO <sub>4</sub> + Ethylenediamide; final leaching in 0.5 M H <sub>2</sub> SO <sub>4</sub>
28	GGL-HT1000Ar-H <sub>2</sub> SO <sub>4</sub>	197	0.5	0.02	0.03	1.50	<0.01	<0.01	<0.01	GGL: Mixture of 0.49:0.49:0.02 mol Glucose:Glycin:Fe-lactate; leaching in 0.5 M H <sub>2</sub> SO <sub>4</sub>
29	(FePhen <sub>3</sub> + BP2000)-HT900Ar	174	1.2	0.51	-	-	0.03	-	-	Complexation of phanthroline with hydrated Fe-sulfate

both, RDE as well as fuel cell measurements (columns G and H).

Turnover frequency TOF (in electrons site<sup>-1</sup>·s<sup>-1</sup>) :

$$TOF = \frac{J_{K(A/g)}}{S_D(\text{sites/g})} \cdot \frac{1}{e}, \quad (5)$$

In this equation, e stands for the elementary charge of the electron.

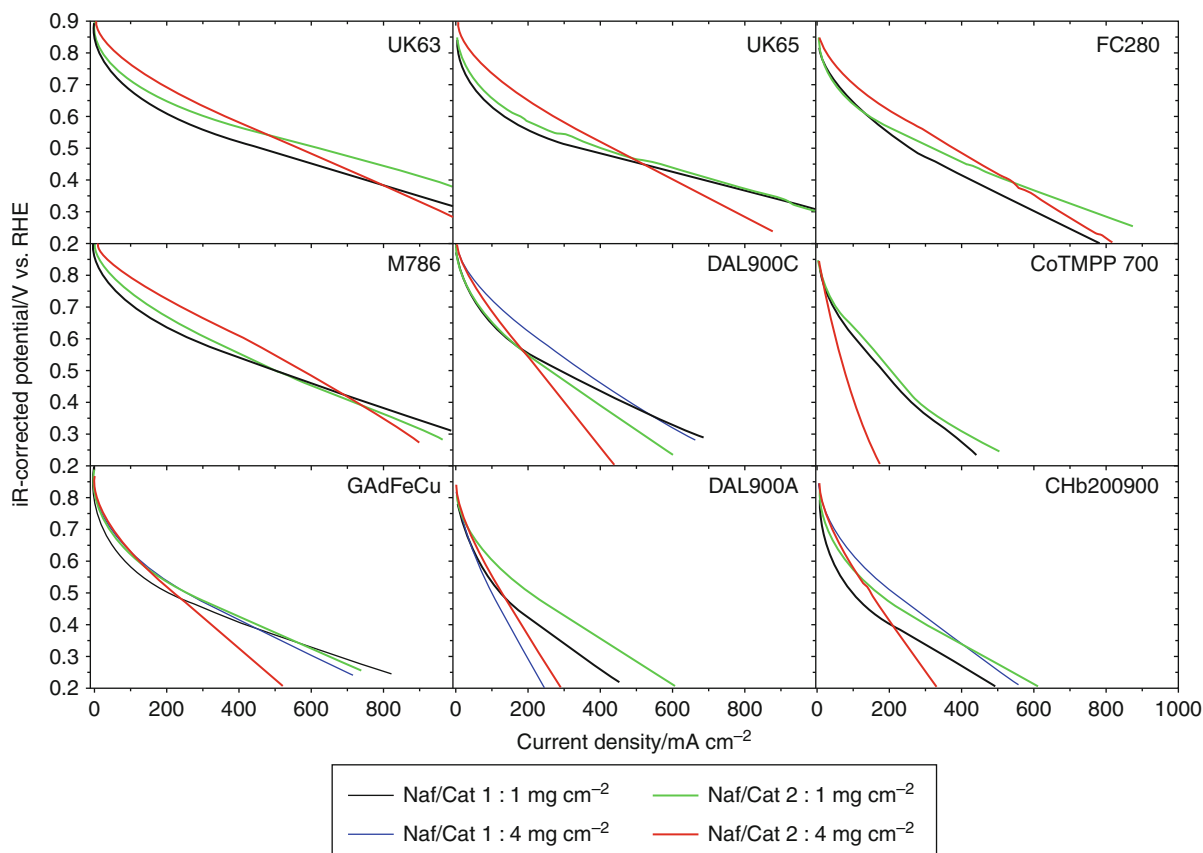
Fuel cell measurements excellently reflect the actual performance of a catalyst and its applicability. Besides the kinetic behavior, however, the mass transport properties (proton, electron, and oxygen transport as well as water removal) and the internal resistance affect the performance in fuel cells. Thus, it becomes difficult to estimate the kinetic properties of a catalyst just from FC measurements. Rotating (Ring) Disk Electrode (R(R)DE) measurements more precisely reflect the kinetic properties (ORR activity and selectivity). Therefore, in this work also, R(R)DE measurements were taken into account. A comparison of publications showed that measurement conditions (for PEM-FC

and R(R)DE) often vary considerably from one laboratory to another, which hinders a direct comparison of the different materials [7, 114, 115, 160, 165, 166, 168–171, 173, 175, 177, 196–206]. In order to enable a better evaluation, F. Jaouen initiated a cross-laboratory comparison of FC and RDE measurements for various Me-N-C catalysts [160]. In Fig. 20, fuel cell measurements of these catalysts are shown.

It was found that the catalysts from different institutes behaved completely different when changing, e.g., the Nafion content or the catalyst loading in FC and RDE tests, as shown for FC measurements in Fig. 20.

Regarding the volumetric current density, best Fe-N-C catalysts yield about 25–40% of the DOE target of 325 A/cm<sup>3</sup> envisaged for 2015 (see lines 3–6, in Table 2) [165, 207, 208]. P. Zelenay's working group published for one of their catalysts a volumetric activity of 40% of the DOE target for 2015 (line 3) [208]. In that preparation route, catalysts were prepared from a mixture of cyanamide, iron sulfate, and sucrose. Further details of the preparation procedure are not yet published. The activity of the so-called PANI-based





**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.** Figure 20 Fuel cell tests of NNMC prepared by different preparation methods (for catalyst's preparation routes, compare Table 2 and [160]). The given FC measurements were made with different Nafion-to-catalyst ratios (Naf/Cat). The figure was taken from [160]; reproduced with permission of the American Chemical Society

catalysts of the same group was boosted by optimization of the preparation steps and the addition of ethylenediamine [189, 207, 208]. Comparing the results of lines 6 and 14, it becomes apparent that by the optimization, the volumetric activity was increased by a factor of 18.

The catalysts in lines 4–5 have been produced via the pore-filling method (PFM) [165]. The preparation is described in the section “Alternative Center Generation During Heat Treatment.” A catalyst produced in this manner achieved about 50% of the electron transfer rate of commercial platinum. Similar catalysts, but without ball milling, were already described by Bron et al. in 2002 [174, 178, 188]. The comparison clearly shows that the ball milling has a considerable influence on the achievable activity (lines 4, 5, and 29).

Comparing the volumetric current densities in column I related to the catalyst manufacturing method, it becomes apparent that all catalysts in the upper third of Table 2 were produced in a multistep preparation and/or optimization process. Furthermore, in nearly all procedures, one preparation step is performed in ammonia at  $T \geq 800^\circ\text{C}$ . In 2008, different groups have shown that by performing a second pyrolysis step in ammonia, the kinetic current density can be enhanced drastically [162, 166, 167], a less pronounced improvement was found when other gases (like  $\text{CO}_2$ ,  $\text{N}_2$ ) were used [112, 162]. If previous to the  $\text{NH}_3$ -treatment a high-energy ball milling is performed, an up to 30-fold increase in current density was achieved as can be learned from columns 8 and 20 in Table 2. The ball milling itself does not affect the kinetics of the

catalyst; however, it can lead to enhanced mass transport properties [209]. Induced by the ball milling, the carbon agglomerates formed during this template-assisted preparation were cracked down allowing subsequently the  $\text{NH}_3$  to react more efficiently with the catalyst (compare lines 8 and 9) [209]. Induced by the burn-off, higher site densities are obtained, and a formation of new and/or different catalytic centers might appear [63, 129, 167, 170]. On the other hand, the turnover frequency of the catalysts is improved. To illustrate this behavior, compare lines 4, 5, 12; lines 17, 19; lines 8, 20 and lines 9, 23. Possible reasons for the enhanced kinetics could be (1) the formation of additional, maybe different catalytic centers (i.e.,  $\text{FeN}_4$  vs.  $\text{FeN}_{2+2}$  in micropores, compare section “Molecular Centers in Carbonized Materials,” Figs. 16 and 19), (2) a general change of the reduction mechanism, or (3) an improvement of the carbon matrix with respect to its electron donor properties [112, 136, 167, 191].

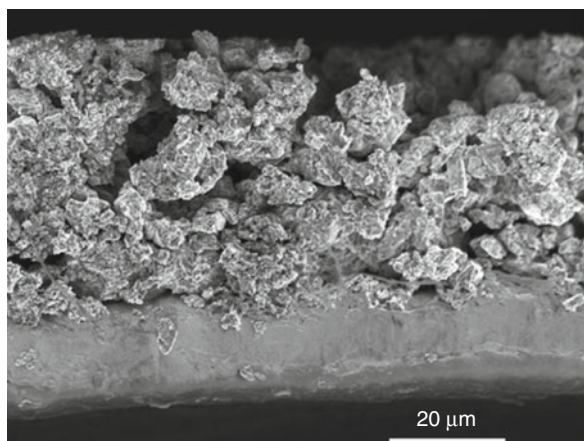
As described in section “Alternative Center Generation During Heat Treatment,” some preparation approaches failed by the attempt to increase the site density above a certain critical value [170, 172]. In Table 2, column C, the site densities are listed. It can be deduced that catalysts prepared by the use of a carbon support usually reveal smaller site densities compared to those where exclusively organic precursor molecules were used as carbon sources. In template-assisted preparation processes, an acid leaching (removal of the template) was performed after a first pyrolysis and before the catalyst was further processed [39, 115, 130–133, 159, 160, 162, 177, 186, 189, 197, 201, 209]. The obtained catalysts exhibit site densities of  $3 \cdot 10^{20}$  up to  $7 \cdot 10^{20}$  centers per gram (compare Table 2). In contrast, the utilization of carbon blacks as support often leads to site densities  $< 1.5 \cdot 10^{20}$  centers per gram. In these catalysts, the number of active sites should be further increased for FC application.

When high current densities are requested, a lack in the presence of one (or more) of the reactants at the three phase boundary often limits the performance. Another issue concerns the removal of reaction water out of the micro- and mesopore structure which is lowering the kinetics of reactive centers. Thus, in order to optimize the mass transport properties, either the electron or proton conductivity has to be augmented or the pore structure of the catalyst has to be

improved to enable better oxygen diffusion to the active centers. Furthermore, the catalyst should be sufficiently hydrophobic in order to guarantee a fast removal of the water to prevent any blocking of active sites. Preferentially, surface groups on the carbon increase the hydrophilicity of the catalyst. Especially disordered carbon can provide higher concentrations of such groups caused by a high number of defects in comparison to well-ordered carbon. Thus, in turn, a high degree of graphitization is beneficial for the water transport properties [207]. As discussed in the next section, graphitic carbon can also enhance the long-term stability.

When switching from RDE to fuel cell measurements, an increase of catalytic activity is expected, simply because of the higher operation temperature. Column F lists the ratio of achieved activities in fuel cell and RDE measurements ( $J_{\text{FC}}/J_{\text{RDE}}$ ) at  $U = 0.8$  V. Looking at catalysts, which were produced without commercial carbon support, difficulties in the fuel cell application (as expressed by a low current ratio FC/RDE in column F) can be recognized with the exception of the catalyst in line 8 of Table 2. This catalyst is similar to the one in line 9 except of a ball milling previous to the ammonia treatment. Therefore, a main reason might be the presence of substantially larger carbon agglomerates, which inhibit a homogeneous preparation of the gas diffusion electrode (GDE) [160, 162, 209], as shown in Fig. 21.

Looking at the template-assisted preparation techniques, the physicochemical properties of the in-situ formed carbon could be of lower quality compared to commercial carbon supports. For fuel cell application, an optimization with respect to (1) the electrical conductivity, (2) the agglomerate or particle size, and (3) the transport properties (hydrophilicity) is necessary. However, optimal physicochemical properties of the carbon might not be obtained at the maximum of volumetric current density (determined by  $S_{\text{D}}$  and TOF). Therefore, approaches to improve catalysts with an in-situ formed carbon should be also done under utilization of FC measurements in order to evaluate the mass transport kinetics. Nevertheless, optimization remains challenging as a good balance between carbon properties and ORR performance must be found. Not-optimized carbon could not only lower the achievable activity, it could also affect the long-term stability as will be discussed in the following section.



**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.**  
**Figure 21**

GDE of a catalyst prepared by the oxalate-supported pyrolysis. The figure was taken from [162], reproduced by permission of ECS – The Electrochemical Society

### Stability

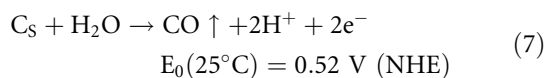
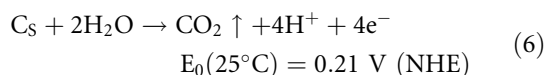
In this part, possible degradation mechanisms and the currently most promising catalysts with respect to long-term stability will be discussed. For the final implementation, the observed stability is crucial. To optimize the material system, however, it has to be rationalized which processes lead to a decrease of ORR activity.

Regarding non-pyrolyzed macrocycles, hydrogen peroxide formed in the reduction cycle can lead to broken bonds between the substituents and the tetrapyrrole core. Changes in the electronic structure connected therewith cause a decrease of the reduction activity. In acidic conditions as given in PEM-FC, demetallation can also cause an activity decline [71, 87, 98, 102, 210, 211]. Furthermore, carbon monoxide is able to bind irreversibly to the metal centers and thereby deactivating them [212]. With respect to Eq. 7, carbon monoxide might be formed by carbon oxidation during FC operation.

Better stabilities were achieved investigating heat-treated catalysts (section “Molecular Centers in Carbonized Materials”). Nevertheless, in most of the cases, even after the heat treatment, considerable decreases of performance were already observed after a few hours of operation time [69, 159, 168, 184, 188, 213–216]. For catalysts prepared by a heat treatment of either iron

porphyrin or iron phthalocyanine, it was shown that they are tolerant toward CO [217, 218]. Therefore, we assume that in general, CO-poisoning can be ruled out as degradation mechanism for carbonized materials. Possible degradation mechanisms could be addressed to (1) a corrosion of carbon as known from platinum catalysts, (2) an inactivation by leaching of active sites, or (3) a deactivation of active species (e.g., by blocking of the centers by intermediates or the final products) [71, 98, 99, 102, 190, 191, 210, 211, 219, 220].

The thermodynamic process of carbon oxidation already starts at potentials  $>0.21$  V. For Pt/C catalysts, SEM cross-sectional images illustrated a significant decrease of the carbon layer thickness, causing a detachment of the platinum particles or an increase of the particle size [8, 110, 220]. In Eqs. 6 and 7, the oxidation reactions of carbon to carbon dioxide (Eq. 6) and carbon monoxide (Eq. 7) are given.



Both reactions are initiated by the presence of water. This fact again underlines the importance of an optimized water management. Independent of the nature of the active centers in Me-N-C catalysts, they are presumably integrated into a carbon matrix, either as nitrogen heteroatoms (Fig. 8), MeN<sub>4</sub>-centers (Fig. 16), or as MeN<sub>2+2</sub>-centers in micropores between two adjacent graphene layers (Fig. 19). As discussed above, in N<sub>4</sub>-macrocycles, a decrease of the conjugated  $\pi$ -electron system causes a lowering of the ORR activity [41, 42, 80, 81, 221]. It can be assumed that a carbon burn-off in heat-treated materials can cause a similar effect (see below, discussion related to Fig. 25).

Herranz et al. showed that FeN<sub>2+2</sub>-centers have two different states of catalytic activity depending on whether an NH<sup>+</sup>-group or an NH-anion-group is present in its environment [191]. For the first case, very high turnover frequencies are reported. The anion blocking causes a drastic decrease of the turnover frequency and will appear during RDE or FC tests. Nevertheless, it was found that this deactivation process is reversible and that the centers can be activated by performing a thermal or chemical recover

treatment [191]. As such treatments are not applicable during FC operation, however, the effect in general should be inhibited if it is the main reason for activity decay of a specific catalyst.

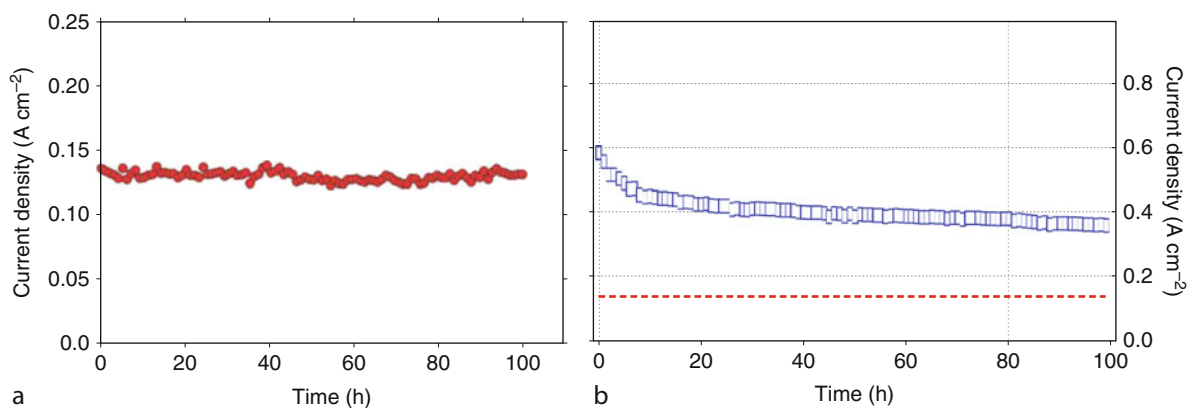
In 2006, Bashyam and Zelenay published a stable performance of a catalyst entirely produced without heat treatment obtaining stability for a runtime of 100 h. The performance of the catalyst is shown in Fig. 22a. It was prepared by polymerization of pyrrole in the presence of a carbon-supported cobalt salt. Caused by the polymerization, an extended  $\pi$ -electron system was formed leading to good electron conductivity. At that time, the authors suggested that cobalt ions – twofold coordinated by nitrogen atoms – are responsible for the activity [177].

However, the disadvantage of this material was a low catalytic activity (23 A/g at 0.4 V at a catalyst loading of 6 mg/cm<sup>2</sup>). Recent approaches of the same group showed that besides pyrrole, also other nitrogen heterocycles and metal salts can be used as nitrogen and metal sources, respectively [189, 208]. Similar to non-pyrrolyzed macrocycles, the activity can be enhanced by a heat treatment [177, 189]. A comparison of different nitrogen heterocycles used in the preparation showed that the selection of the polymer significantly affects activity and stability – even after a heat treatment. The Fe/PANI/C-catalyst supported on multiwall carbon

nanotubes (MWCNT) achieved an activity of 75 A/g (at 0.4 V) and showed no degradation within the investigated time of 500 h [186, 189]. Therefore, this catalyst is currently the most stable Fe-N-C catalyst, exhibiting even a good catalytic activity (remark: considering catalysts, which have been investigated in fuel cells with respect to their stability). The preparation is similar to that of a pyrrole-based catalyst.

The currently most active catalysts prepared by the Pore-Filling-Method (PFM) from J.-P. Dodelet's group show an average degradation of 0.38%/h for 100 h potentiostatic operation with respect to the initial activity (compare [165, 192]), the related curve is shown in Fig. 22b. The measuring conditions thereby were more or less identical to Bashyam's and Zelenay's experiments. It seems that for the degradation of this PFM-catalyst, one can distinguish an initial fast degradation ( $t < 20$  h) which is superimposed by a weaker and slower one ( $t > 20$  h:  $-0.08\%/h$ ).

In the work of Wu et al., structural changes of a Fe/PANI/KB300 catalyst were investigated before and after a runtime of 500 h [189]. After a long-term test, the percentage of oxygen in the matrix was drastically increased, that the authors attributed to the oxidation of carbon and nitrogen atoms. Particularly striking were a complete loss of sulfur (which was present in the polymerization agent) and a slight decrease of the



**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.** Figure 22 Stability measurements of (a) a Co-PPy-CB and (b) a Fe/CB-PFM catalysts. Both catalysts were measured under nearly the same conditions at 80°C, backpressures of 2 atm, gases at 100% RH with catalyst loadings of 6 mg/cm<sup>2</sup> and 5.6 mg/cm<sup>2</sup> for the Co-PPy-CB and Fe/CB-PFM catalysts, respectively. The figure in (a) was taken from [177]; copyright 2006, reproduced with permission of Nature Publishing Group. The figure in (b) was taken from [165], reprinted with permission from AAAS

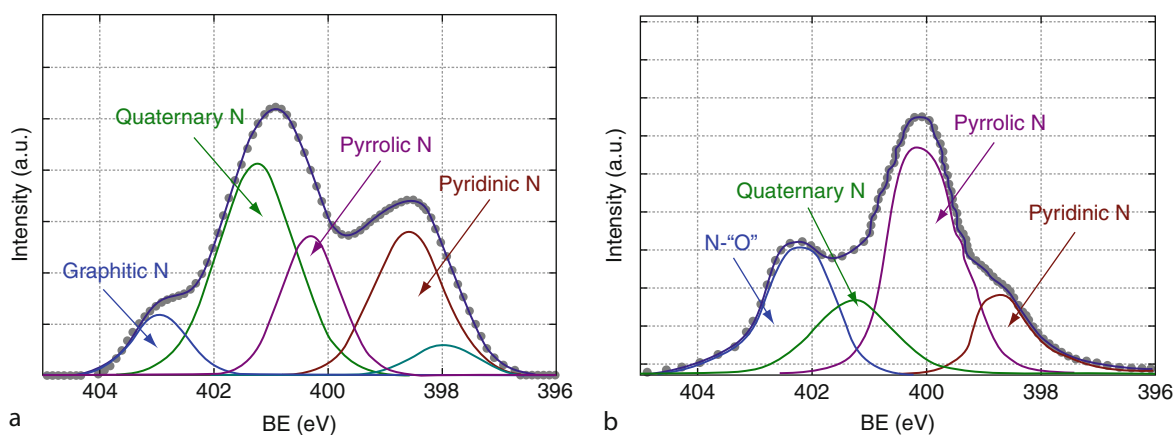
total nitrogen content (from 6.3% to 5.6%). Changes in the relative composition of the N1s spectra were significant. A comparison of the spectra before and after 500 h fuel cell test time is shown in Fig. 23. It is evident that the long-term performance causes an energetic shift of the nitrogen species toward higher bonding energies. This was mostly attributable to the oxidation of pyridinic nitrogen atoms. In the investigated time-frame, the catalyst lost 17% of its initial activity. Similar shifts in the XPS N1s-spectra were observed by Liu et al. and by Kramm [112, 115].

In the case of Liu et al., significantly higher activity losses compared to Wu et al. were observed ( $-0.46\%/h$  vs  $-0.04\%/h$ ), while Kramm did not observe any significant loss (compare Fig. 25). Liu et al. assigned the overall ORR activity to highly active pyridinic and less active graphitic nitrogen atoms (both without iron participation) [115]. They explained their XPS results with a protonation of pyridinic nitrogen atoms in contact with an acidic electrolyte; this process was proposed to be the main reason for the observed deactivation. On the other hand, graphitic nitrogen atoms were assigned to be less active but more stable in acidic environment [115]. Kramm investigated the structural changes of a Fe-N-C catalyst via XPS, NAA and  $^{57}\text{Fe}$  Mößbauer spectroscopy [112]. The catalyst was produced by the sulfur-assisted oxalate-supported pyrolysis of FeTMPPCI; in a subsequent acid leaching step, excess iron was removed. Before the long-term

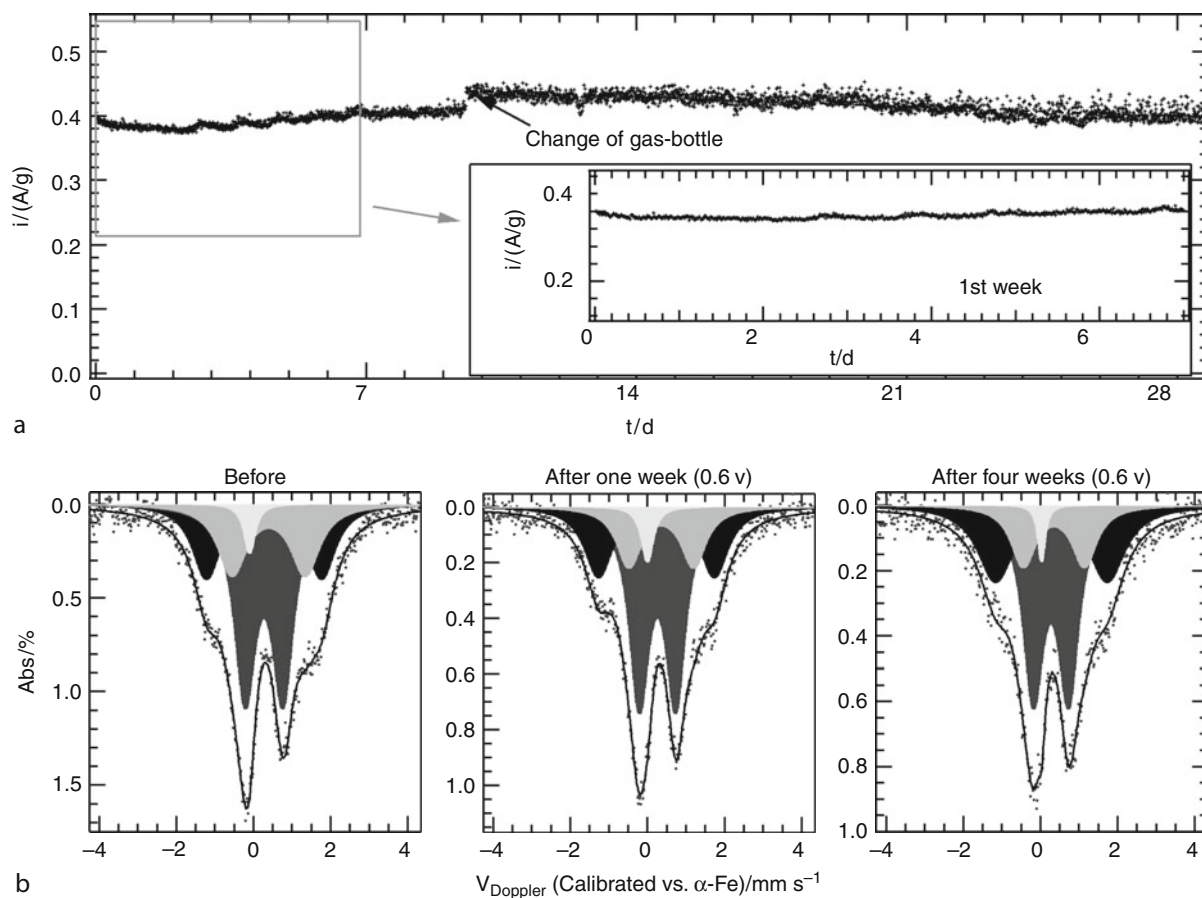
measurement was performed, the electrode was kept in distilled water for 24 h in order to initiate a complete swelling of the PTFE previous to the stability test shown in Fig. 24.

The preconditioned electrode showed no loss in activity over a period of 4 weeks and no significant change of the composition in Mößbauer spectra or iron content. The changes of the N1s spectra, however, were similarly pronounced as described by Wu et al. [112, 189]. A comparison of the three references discussed above leads to the conclusion that pyridinic nitrogen atoms might affect the activity of the catalysts up to a certain degree. However, it has been demonstrated that pyridinic nitrogen atoms cannot explain the catalytic activity alone [112].

In an early work of Faubert et al. [140], the influence of the pyrolysis temperature on the stability of MeTPP/C was investigated (Me = Fe, Co). Figure 25 shows the changes of the current density (related to the surface area and the metal content) as a function of performance time. It is evident that catalysts prepared from FeTPP achieve higher current densities and better stabilities than CoTPP-based catalysts. Apart from a pronounced loss in activity within the first 5 min, the FeTPP/C catalysts, produced at  $800^\circ$  and  $900^\circ$  C, demonstrated a nearly stable performance. At even higher pyrolysis temperatures, the stability performance was similar to that of the platinum catalyst (2 wt% Pt/C), which was measured as a reference.



**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.** Figure 23 N 1s XP spectra of a Fe-PANI-C catalyst (a) before and (b) after a 500-h long-term stability test. The figure was taken from [189]; reproduced by permission of ECS – The Electrochemical Society



**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.** Figure 24 Potentiostatic ( $U = 0.6$  V) long-term stability test in  $O_2$ -flushed 0.5 M  $H_2SO_4$  of a Fe-N-C-catalyst at RT (a). The insert gives a magnification of the first week of measuring time. The catalyst was prepared by the sulfur-assisted oxalate-supported pyrolysis of FeTMPPCl (800°C,  $N_2$ ), with acid leaching. The structural features related to iron were studied by Mössbauer spectroscopy at RT; (b) before, after 1 week and after 4 weeks of measuring time. The figure is adapted from Figures 7-7 and 7-8 of [112]; copyright 2009 by Südwestdeutscher Verlag für Hochschulschriften

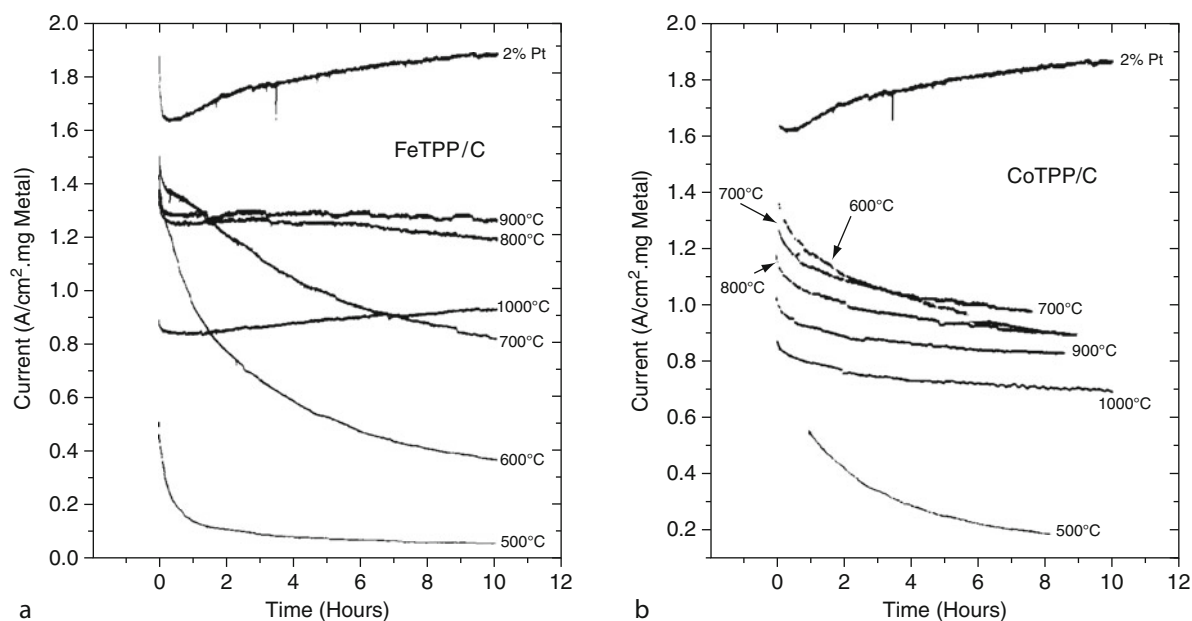
The authors assigned the enhanced stability to an increased amount of graphitic carbon [140].

Recent results published by the same group [222, 223] and the comparison of various carbon-supported Fe/PANI/C catalysts confirm this observation [189]. One might summarize that in fuel cells, obviously the long-term stability seems to depend on the corrosion resistivity of the carbon matrix for both, Fe-N-C and Pt/C catalysts [8, 224].

In order to avoid any activity loss, the catalysts must be stabilized: For Pt/C catalysts, it was shown that the implementation of metal oxide particles into the carbon matrix can enhance the corrosion resistivity

substantially. Furthermore, as far as possible, the catalysts should comprise fractions of graphitic carbon which prevent a fast oxidation.

It is believed that the search for alternative supports for Fe-N-C catalysts will become dominant in the future. Especially, different transition metal oxynitrides ( $TiO_xN_y$ ,  $TaO_xN_y$ ,  $ZrO_xN_y$ ), and Nb- or Sb-doped  $TiO_2$  or  $SnO_2$  or PANI appear to be promising in this context [55, 224–227]. It will remain a major challenge to enable the implementation of molecular centers onto the surface of the new support material. It has already been pointed out that PANI possesses a considerable potential with this respect [189]. Since none of the



**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction.** Figure 25 Stability tests of carbon-supported FeTPP (a) and CoTPP (b), both pyrolyzed at various temperatures as indicated in the graphs. The figures were taken from [140], reproduced with permission of Elsevier

other promising support materials contain carbon, it is questionable how catalytically active Fe-N-C catalysts could be implemented.

With this respect, the work from Atanasoski and coworkers is promising (compare section “Transition Metal Carbides, Nitrides and Chalcogenides”) [35]. By performing a heat treatment of their sputtered C-N<sub>x</sub>:Fe films, the activity was drastically enhanced but still much lower compared to macrocycle-based catalysts. However, when titanium carbide was used as support instead of carbon, a high stability was obtained. The fact that by changing the support, an essentially better durability was obtained is an important result as it shows that even for catalysts based on molecular centers, alternative support materials can be utilized and that the interaction between the support and the catalytic centers might be crucial for the optimization of those catalysts for a fuel cell application.

However, it has to be kept in mind that a major challenge of these catalysts is to improve their transport properties which are by far not ideal as can be recognized from Tafelplots deduced from RDE measurements. To investigate the water management in platinum-based fuel cells, the use of neutron scattering

has meanwhile proven to be an effective tool [228, 229]. Such investigations could also play a major role for further optimization of Fe-N-C catalysts.

### Future Direction

Comparing the different types of NNMC, the Fe-N-C catalysts are presently the most promising candidates to replace platinum or other noble metal catalysts. These materials exhibit the highest activities and catalyze with high selectivity the direct reduction of oxygen to water. Since 2009, a significant improvement in comparison to previous results was made; here especially the work of Dodelet and coworkers and Zelenay and coworkers should be mentioned [165, 189]. In general, stable fuel cell performance over periods of up to 1000 h has already been demonstrated (Table 3). Although long-term stability and high catalytic activity appear to be difficult to combine into one material, based on today’s knowledge of the structural characteristics that affect activity and/or stability, further improvements of FC performance are expected. One major step to an improved stability may be through alternative support materials.

**Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction. Table 3**  
Comparison of different Fe–N–C catalysts with respect to their stability

	A	B	C	D	E	F	G	H
	Catalyst	Lit.	Conditions	Comm.	mW/cm <sup>2</sup>		time/ h	Performance (%/h)
					Start	End		
1	(FeTMPPCl/non- $\mu$ pCB)–HT950Ar, 5.2 wt%Fe	222	FC80, H <sub>2</sub> /O <sub>2</sub> : 1 bar, 2 mg/cm <sup>2</sup>	0.6	2.58	6.24	15	9.46
2	(FeTMPPCl/non- $\mu$ pCB)–HT950Ar–H <sub>2</sub> SO <sub>4</sub> , 5.2 wt%	222	FC80, H <sub>2</sub> /O <sub>2</sub> : 1 bar, 2 mg/cm <sup>2</sup>	0.6	5.1	6.6	15	1.96
3	PANI–Fe–MWCNT (HT900, N <sub>2</sub> –H <sub>2</sub> SO <sub>4</sub> )	189	FC80, H <sub>2</sub> /Air: 2 bar, 4 mg/cm <sup>2</sup>	0.4	120	120	500	0.00
4	PPy–Co–C	177	FC80, H <sub>2</sub> /Air: 2 bar, 4 mg/cm <sup>2</sup>	0.4	56	56	100	0.00
5	PPy–Co–C	177	FC80, H <sub>2</sub> /O <sub>2</sub> : 2 bar, 6 mg/cm <sup>2</sup>	0.7	24.5	24.5	50	0.00
6	Fe/Fe/S–HT800N <sub>2</sub> –HCl, 24h_swelling_inH <sub>2</sub> O	112	T = 25°C, 0.5 M H <sub>2</sub> SO <sub>4</sub> , O <sub>2</sub> , 4 mg/cm <sup>2</sup>	0.6	0.96	0.96	672	0.00
7	3M–NiANI–HT_NH <sub>3</sub> *	166	FC80, H <sub>2</sub> /O <sub>2</sub> : 3/4.3 bar	0.6	0.0024	0.0024	0.5	0.00
8	PANI–Fe <sub>3</sub> Co–C (HT900, N <sub>2</sub> –H <sub>2</sub> SO <sub>4</sub> )	189	FC80, H <sub>2</sub> /Air: 2 bar, 4 mg/cm <sup>2</sup>	0.4	148	132	600	–0.02
9	PANI–Fe <sub>3</sub> Co–C (HT900, N <sub>2</sub> –H <sub>2</sub> SO <sub>4</sub> )	186	FC80, H <sub>2</sub> /Air: 2 bar, 4 mg/cm <sup>2</sup>	0.4	140	130	400	–0.02
10	PANI–Fe–C (HT900, N <sub>2</sub> –H <sub>2</sub> SO <sub>4</sub> )	189	FC80, H <sub>2</sub> /Air: 2 bar, 4 mg/cm <sup>2</sup>	0.4	120	100	500	–0.03
11	((CB–HCl–HNO <sub>3</sub> ) + Co/Fe/N) – HT900Ar–H <sub>2</sub> SO <sub>4</sub>	119	FC80, H <sub>2</sub> /O <sub>2</sub> , no backpressure	200 mA/cm <sup>2</sup>	70	60	450	–0.03
12	PANI–Fe–KB300 (HT900, N <sub>2</sub> –H <sub>2</sub> SO <sub>4</sub> )	189	FC80, H <sub>2</sub> /Air: 2 bar, 4 mg/cm <sup>2</sup>	0.4	116	90	500	–0.04
13	PANI–Co–C (HT900, N <sub>2</sub> –H <sub>2</sub> SO <sub>4</sub> )	189	FC80, H <sub>2</sub> /Air: 2 bar, 4 mg/cm <sup>2</sup>	0.4	100	90	180	–0.06
14	Fe/Fe/S–HT800N <sub>2</sub> –HCl	112	GDE, T = 25°C, O <sub>2</sub> , 4 mg/cm <sup>2</sup>	0.6	0.96	0.7032	168	–0.16
15	(Fe,Co)/EDA + SiO <sub>2</sub> –HT1000Ar–NaOH–H <sub>2</sub> SO <sub>4</sub> –HT1000Ar	115	FC75, H <sub>2</sub> /Air: 2 bar, 2 mg/cm <sup>2</sup>	0.4	140	110	100	–0.21
16	CHb200900	204	FC80, H <sub>2</sub> /Air:50/250, 10 mg/cm <sup>2</sup>	0.5	–	–	200	–0.22
17	CHb350900C	204	FC80, H <sub>2</sub> /Air:50/250, 10 mg/cm <sup>2</sup>	0.5	–	–	200	–0.33
18	PFM1: Fe/Phen/BP–HT1050Ar–HT950NH <sub>3</sub>	165	FC80, H <sub>2</sub> /Air: 2 bar, 5.6 mg/cm <sup>2</sup>	0.4	240	150	100	–0.38
19	(Fe,Co)/EDA + SiO <sub>2</sub> –HT800Ar–NaOH–H <sub>2</sub> SO <sub>4</sub> –HT800Ar	115	FC75, H <sub>2</sub> /Air: 2 bar, 2 mg/cm <sup>2</sup>	0.4	204	120	90	–0.46
20	PPy–Fe–C (HT900, N <sub>2</sub> –H <sub>2</sub> SO <sub>4</sub> )	189	FC80, H <sub>2</sub> /Air: 2 bar, 4 mg/cm <sup>2</sup>	0.4	196	100	100	–0.49



Polymer Electrolyte Membrane Fuel Cells (PEM-FC) and Non-noble Metal Catalysts for Oxygen Reduction. Table 3 (Continued)

	A	B	C	D	E		G	H
	Catalyst	Lit.	Conditions	Comm.	mW/cm <sup>2</sup> Start End		time/ h	Performance (%/h)
21	PFM1: Fe/Phen/BP-HT1050Ar-HT950NH <sub>3</sub>	165	FC80, H <sub>2</sub> /O <sub>2</sub> : 2 bar, 5.6 mg/cm <sup>2</sup>	0.5	385	165	100	-0.57
22	(FePhen3 + BP2000)-HT900Ar	178	GDE, T = 25°C, 0.5 M H <sub>2</sub> SO <sub>4</sub> , Air	?	13.5	3	40	-1.94
23	(Graphite + BM5)-FeAc-HT950NH <sub>3</sub>	184	FC80, H <sub>2</sub> /O <sub>2</sub> : 4/2 bar, 1.1 mg/cm <sup>2</sup>	0.3	135	69	24	-2.04
24	(Graphite + BM0.5)-FeAc-HT950NH <sub>3</sub>	184	FC80, H <sub>2</sub> /O <sub>2</sub> : 4/2 bar, 1.1 mg/cm <sup>2</sup>	0.3	102	37.5	24	-2.63
25	(Graphite + BM10)-FeAc-HT950NH <sub>3</sub>	184	FC80, H <sub>2</sub> /O <sub>2</sub> : 4/2 bar, 1.1 mg/cm <sup>2</sup>	0.3	43.5	14.4	24	-2.79
26	(FeTMPPCl/non-μpCB)-HT950NH <sub>3</sub> , 0.4 wt%Fe	222	FC80, H <sub>2</sub> /O <sub>2</sub> : 1 bar, 2 mg/cm <sup>2</sup>	0.6	108	13.2	15	-5.85
27	(FeTMPPCl/non-μpCB)-HT950Ar, 0.4 wt% Fe	222	FC80, H <sub>2</sub> /O <sub>2</sub> : 1 bar, 2 mg/cm <sup>2</sup>	0.6	6	0.66	15	-5.93

Apart from this “hot” topic, one should keep in mind that catalysts based on transition metal chalcogenides, nitrides, and carbides also have the potential to be used without carbon support. Considering the tremendous carbon corrosion problems especially at higher temperature (>110°C), carbon-free electrode structures will be an important aspect of future research.

## Bibliography

- Markovic NM, Radmilovic V, Ross PN (2003) Catalysis and electrocatalysis at nanoparticle surfaces. Marcel Dekker, New York/Basel
- Mukerjee S, Srinivasan S (1993) Enhanced electrocatalysis of oxygen reduction on platinum alloys in proton exchange membrane fuel cells. *J Electroanal Chem* 357:201–224
- Zhang J, Lima FHB, Shao MH, Sasaki K, Wang JX, Hanson J, Adzic RR (2005) Platinum monolayer on non-noble metal-noble metal core-shell nanoparticle electrocatalysts for O<sub>2</sub> reduction. *J Phys Chem B* 109:22701–22704
- Stamenkovic VR, Mun BS, Mayrhofer KJJ, Ross PN, Markovic NM (2006) Effect of surface composition on electronic structure, stability and electrocatalytic properties of Pt-transition metal alloys: Pt-skin versus Pt-skeleton surfaces. *J Am Chem Soc* 128:8813–8819
- Strasser P, Koh S, Anniyev T, Greeley J, More K, Yu C, Liu Z, Kaya S, Nordlund D, Ogasawara H, Toney MF, Nilsson A (2010) Lattice-strain control of the activity in dealloyed core-shell fuel cell catalysts. *Nat Chem* 2:454–460
- Adzic RR, Zhang J, Sasaki K, Vukmirovic MB, Shao M, Wang JX, Nilekar AU, Mavrikakis M, Valerio JA, Uribe F (2007) Platinum monolayer fuel cell electrocatalysts. *Top Catal* 46:249–262
- Zhang S, Yuan XZ, Hin JC, Wang H, Friedrich KA, Schulze M (2009) A review of platinum-based catalyst layer degradation in proton exchange membrane fuel cells. *J Power Sources* 194:588–600
- Mayrhofer KJJ, Meier JC, Ashton SJ, Wiberg GKH, Kraus F, Hanzlik M, Arenz M (2008) Fuel cell catalyst degradation on the nanoscale. *Electrochem Commun* 10:1144–1147
- Delacote C, Bonakdarpour A, Johnston CM, Zelenay P, Wieckowski A (2008) Aqueous-based synthesis of ruthenium-selenium catalyst for oxygen reduction reaction. *Faraday Discuss* 140:269–281
- Fiechter S, Dorbandt I, Bogdanoff P, Zehl G, Schulenburg H, Tributsch H, Bron M, Radnik J, Fieber-Erdmann M (2007) Surface modified ruthenium nanoparticles: structural investigation and surface analysis of a novel catalyst for oxygen reduction. *J Phys Chem C* 111:477–487
- Guinel MJF, Bonakdarpour A, Wang B, Babu PK, Ernst F, Ramaswamy N, Mukerjee S, Wieckowski A (2009) Carbon-supported, selenium-modified ruthenium-molybdenum catalysts for oxygen reduction in acidic media. *ChemSusChem* 2:658–664
- Haas S, Hoell A, Zehl G, Dorbandt I, Bogdanoff P, Fiechter S (2008) Structural investigation of carbon supported Ru-Se based catalysts using anomalous small angle X-ray scattering. *ECS Trans* 6:127–138
- Loponov KN, Kriventsov VV, Nagabhushana KS, Boennemann H, Kochubey DI, Savinova ER (2009) Combined in situ EXAFS

- and electrochemical investigation of the oxygen reduction reaction on unmodified and Se-modified Ru/C. *Catal Today* 147:260–269
14. Zehl G, Schmithals G, Hoell A, Haas S, Hartnig C, Dorbandt I, Bogdanoff P, Fiechter S (2007) On the structure of carbon-supported selenium-modified ruthenium nanoparticles as electrocatalysts for oxygen reduction in fuel cells. *Angew Chem Int Ed* 46:7311–7314
  15. Cremers C, Scholz M, Seliger W, Racz A, Knechtel W, Rittmayr J, Grafwallner F, Peller H, Stimming U (2007) Developments for improved direct methanol fuel cell stacks for portable power. *Fuel Cells* 7:21–31
  16. Garsuch A, Michaud X, Böhme K, Wagner G, Dahn JR (2009) Fuel cell performance of templated Ru/Se/C-based catalysts. *J Power Sources* 189:1008–1011
  17. Whipple DT, Jayashree RS, Egas D, Alonso-Vante N, Kenis PJA (2009) Ruthenium cluster-like chalcogenide as a methanol tolerant cathode catalyst in air-breathing laminar flow fuel cells. *Electrochim Acta* 54:4384–4388
  18. Wippermann K, Richter B, Klafki K, Mergel J, Zehl G, Dorbandt I, Bogdanoff P, Fiechter S, Kaytakoglu S (2007) Carbon supported Ru – Se as methanol tolerant catalysts for DMFC cathodes part II: preparation and characterization of MEAs. *J Appl Electrochem* 37:1399–1411
  19. Lee J-W, Popov BN (2007) Ruthenium-based electrocatalysts for oxygen reduction reaction – a review. *J Solid State Electrochem* 11:1355–1364
  20. Zhang L, Zhang J, Wilkinson DP, Wang H (2006) Progress in preparation of non-noble electrocatalysts for PEM fuel cell reactions. *J Power Sources* 156:171–182
  21. Bennett LH, Cuthill JR, McAlister AJ, Erickson NE, Watson RE (1974) Electronic structure and catalytic behavior of tungsten carbide. *Science* 184:563–565
  22. Colton RJ, Huang JJ, Rabalais JW (1975) Electronic structure of tungsten carbide and its catalytic behavior. *Chem Phys Lett* 34:337–339
  23. Levy RB, Boudart M (1973) Platinum-like behavior of tungsten carbide in surface catalysis. *Science* 181:547–549
  24. Ross PN Jr, Stonehart P (1975) Surface characterization of catalytically active tungsten carbide (WC). *J Catal* 39: 298–301
  25. Binder H, Köhling A, Kuhn W, Lindner W, Sandstede G (1969) Tungsten carbide electrodes for fuel cells with acid electrolyte. *Nature* 224:1299–1300
  26. Bronoel G, Museux E, Leclercq G, Leclercq L, Tassin N (1991) Study of hydrogen oxidation on carbides. *Electrochim Acta* 36:1543–1547
  27. Machida K, Enyo M (1990) Preparation of WC<sub>x</sub> thin films by RF sputtering and their electrocatalytic property for anodic methanol oxidation. *J Electrochem Soc* 137:871–876
  28. Sokolsky DV, Palanker VS, Baybatyrov EN (1975) Electrochemical hydrogen reactions on tungsten carbide. *Electrochim Acta* 20:71–77
  29. Mazza F, Trassatti S (1963) Transition metal oxides as DMFC cathodes without platinum. *J Electrochem Soc* 110:847–849
  30. Lee K, Ishihara A, Mitsushima S, Kamiya N, Ota K (2004) Stability and electrocatalytic activity for oxygen reduction in WC + Ta catalyst. *Electrochim Acta* 49:3479–3485
  31. Bhattarai J, Akiyama E, Habazaki H, Kawashima A, Asami K, Hashimoto K (1998) The passivation behavior of sputter-deposited W-Ta alloys in 12 M HCl. *Corros Sci* 40:757–179
  32. Chou WJ, Yu GP, Huang JH (2003) Corrosion resistance of ZrN films on AISI 304 stainless steel substrate. *Surf Coat Technol* 167:59–67
  33. Zhong H, Zhang H, Liang Y, Zhang J, Wang M, Wang X (2007) A novel non-noble electrocatalyst for oxygen reduction in proton exchange membrane fuel cells. *J Power Sources* 164:572–577
  34. Zhong H, Zhang H, Liu G, Liang Y, Hu J, Yi B (2006) A novel non-noble electrocatalyst for PEM fuel cell based on molybdenum nitride. *Electrochem Comm* 8:707–712
  35. Atanasoski R (2007) Novel approach to non-precious metal catalysts. DOE hydrogen program FY2007 annual progress report 820–824
  36. O'Neill DG, Atanasoski R, Schmoekel AK, Vernstrom GD, O'Brien DP, Jain M, Wood TE (2006) Vacuum deposited non-precious metal catalysts for PEM fuel-cells. *Mater Matters* 1:17–19
  37. Blomquist J, Lang H, Larsson R, Widelöv A (1992) Pyrolysis behaviour of metalloporphyrins: part2: a Mössbauer study of pyrolysed Fe<sup>III</sup> tetraphenylporphyrin chloride. *J Chem Soc Faraday Trans* 88:2007–2011
  38. Bouwkamp-Wijnoltz AL, Visscher W, van Veen JAR, Boellaard E, van der Kraan AM, Tang SC (2002) On active-site heterogeneity in pyrolysed carbon-supported iron porphyrin catalysts for the electrochemical reduction of oxygen: an in situ Mössbauer study. *J Phys Chem B* 106:12993–13001
  39. Koslowski UI, Abs-Wurmbach I, Fiechter S, Bogdanoff P (2008) Nature of the catalytic centres of porphyrin based electrocatalysts for the ORR – A correlation of kinetic current density with the site density of Fe-N<sub>4</sub> centres. *J Phys Chem C* 112:15356–15366
  40. Lefèvre M, Dodelet J-P, Bertrand P (2002) Molecular oxygen reduction in PEM fuel cells: evidence for the simultaneous presence of two active sites in Fe-based catalysts. *J Phys Chem B* 106:8705–8713
  41. van Veen JAR, van Baar JF, Kroese KJ (1981) Effect of heat treatment on the performance of carbon-supported transition-metal chelates in the electrochemical reduction of oxygen. *J Chem Soc Faraday Trans* 77:2827–2843
  42. van Veen JAR, Colijn HA, van Baar JF (1988) On the effect of a heat treatment on the structure of carbon-supported metalloporphyrins and phthalocyanines. *Electrochim Acta* 33:801–804
  43. Bradley Easton E, Bonakdarpour A, Dahn JR (2006) Fe-C-N oxygen reduction catalysts prepared by combinatorial sputter deposition. *Electrochem Sol State Letts* 9:A463–A567
  44. Liu Y, Ishihara A, Mitsushima S, Kamiya N, Ota K (2007) Transition metal oxides as DMFC cathodes without platinum. *J Electrochem Soc* 154:B664–B669

45. Liu Y, Ishihara A, Mitsushima S, Ota K (2010) Influence of sputtering power on oxygen reduction reaction activity of zirconium oxides prepared by radio frequency reactive sputtering. *Electrochim Acta* 55:1239–1244
46. Descorme C, Madier Y, Duprez D (2000) Infrared study of oxygen adsorption and activation on cerium – zirconium mixed oxides. *J Catal* 196:167–173
47. Lisebigler AL, Lu G, Yates JT (1995) Photocatalysis on TiO<sub>2</sub> surfaces: principles, mechanisms, and selected results. *Chem Rev* 95:735–758
48. Witko H, Tokarz-Sobieraj R (2004) Surface oxygen in catalysts based on transition metal oxides: what can we learn from cluster DFT calculations? *Catal Today* 91–92:171–176
49. Kim J-H, Ishihara A, Mitsushima S, Kamiya N, Ota K-I (2007) Catalytic activity of titanium oxide for oxygen reduction reaction as a non-platinum catalyst for PEFC. *Electrochim Acta* 52:2492–2497
50. Mentus SV (2004) Oxygen reduction on anodically formed titanium dioxide. *Electrochim Acta* 50:27–32
51. Kudo A, Miseki YCSR (2009) Heterogeneous photocatalyst materials for water splitting. *Chem Soc Rev* 38:253–278
52. Osterloh FE (2008) Inorganic materials as catalysts for photochemical splitting of water. *Chem Mater* 20:35–54
53. Liu G, Zhang HM, Wang MR, Zhong HX, Chen J (2007) Preparation, characterization of ZrO<sub>x</sub>N<sub>y</sub>/C and its application in PEMFC as an electrocatalyst for oxygen reduction. *J Power Sources* 172:503–510
54. Doi S, Ishihara A, Mitsushima S, Kamiya N, Ota K (2007) Zirconium-based compounds for cathode of polymer electrolyte fuel cell. *J Electrochem Soc* 154:B362–B369
55. Ohgi Y, Ishihara A, Matsuzawa K, Mitsushima S, Ota K (2010) Zirconium oxide-based compound as new cathode without platinum group metals for PEFC. *J Electrochem Soc* 57:B885–B891
56. Clouser SJ, Huang JC, Yeager E (1993) Temperature dependence of the tafel slope for oxygen reduction on platinum in concentrated phosphoric acid. *J Appl Electrochem* 23:597–605
57. Ishihara A, Lee K, Doi S, Mitsushima S, Kamiya N, Hara M, Domen K, Fukuda K, Ota K (2005) Tantalum oxynitride for a novel cathode of PEFC. *Electrochem Solid State Lett* 8:A201–A203
58. Imai H, Matsumoto M, Miyazaki T, Fujieda S, Ishihara A, Tamura M, Ota K (2010) Structural defects working as active oxygen-reduction sites in partially oxidized Ta-carbonitride core-shell particles probed by using surface-sensitive conversion-electron-yield X-ray absorption spectroscopy. *App Phys Lett* 96:191905
59. Ohnishi R, Katayama M, Takanabe K, Kubota J, Domen K (2010) Niobium-based catalysts prepared by reactive radio-frequency magnetron sputtering and arc plasma methods as non-noble metal cathode catalysts for polymer electrolyte fuel cells. *Electrochim Acta* 55:5393–5400
60. Takagaki A, Takahashi Y, Yin F, Takanabe K, Kubota J, Domen K (2009) Highly dispersed niobium catalyst on carbon black by polymerized complex method as PEFC cathode catalyst. *J Electrochem Soc* 156:B811–B815
61. Jasinski R (1964) A new fuel cell cathode catalyst. *Nature (Lond)* 201:1212–1213
62. Jahnke H, Schönborn M, Zimmermann G (1976) Organic dye-stuffs as catalysts for fuel cells. *Top Curr Chem* 61:133–182
63. Gupta SL, Tryk D, Bae I, Aldred W, Yeager EB (1989) Heat-treated polyacrylonitrile-based catalysts for oxygen electroreduction. *J Appl Electrochem* 19:19–27
64. Fierro C, Anderson AB, Scherson DA (1988) Electron donor-acceptor properties of porphyrines, phthalocyanines, and related ring-chelates: a molecular orbital approach. *J Phys Chem* 92:6902–6907
65. Zagal JH (1992) Metallophthalocyanines as catalysts in electrochemical reactions. *Coord Chem Rev* 119:89–136
66. Kobayashi N, Nishiyama Y (1984) Catalytic electroreduction of molecular oxygen at glassy carbon electrodes with immobilized iron porphyrins containing zero, one, or four amino groups. *J Electroanal Chem* 181:107–117
67. van Veen JAR, Visser C (1979) Oxygen reduction on monomeric transition metal phthalocyanines in acid electrolyte. *Electrochim Acta* 24:921–928
68. Melendres CA (1980) Mössbauer and Raman spectra of carbon-supported iron phthalocyanine. *J Phys Chem* 84:1936–1939
69. Bagotzky VS, Tarasevich MR, Radyushkina KA, Levina OA, Andrusyova SI (1977–1978) Electrocatalysis of the oxygen reduction process on metal chelates in acid electrolyte. *J Power Sources* 2:233–240
70. Beck F (1977) The redox mechanism of chelate-catalysed oxygen-cathode. *J Appl Electrochem* 7:239–245
71. Blomquist J, Helgeson U, Moberg LC, Johansson LY, Larsson R (1982) Simultaneous electrochemical and Mössbauer measurements on polymeric iron phthalocyanine oxygen electrodes. *Electrochim Acta* 27:1453–1460
72. Brezina M, Khailil W, Koryta J, Musilová M (1977) Electroreduction of oxygen and hydrogen peroxide catalyzed by hemine and phthalocyanines. *J Electroana Chem* 77:237–244
73. Johansson LY, Mrha J, Larsson R (1973) Elektrokatalyse der O<sub>2</sub>-Reduktion in saurer Lösung mit Hilfe der polymeren Phthalocyanine. *Electrochim Acta* 18:255–258
74. Kazarinov VE, Tarasevich MR, Radyushkina KA, Andreev VN (1979) Some specific features of the metalloporphyrin/electrolyte interface and the kinetics of oxygen electroreduction. *J Electroanal Chem* 100:225–232
75. Larsson R, Mrha J (1973) Katalytische Sauerstoffelektrode in sauren Lösungen. *Electrochim Acta* 18:391–394
76. Maroie S, Savy M, Verbist JJ (1979) ESCA and EPR studies of monomer, dimer, and polymer iron phthalocyanines: involvements for the electrocatalysis of O<sub>2</sub> reduction. *Inorg Chem* 18:2560–2567
77. Zagal JH, Páez M, Tanaka AA, dos Santos JR, Linkous CA (1992) Electrocatalytic activity of metal phthalocyanines for oxygen reduction. *J Electroanal Chem* 339:13–30
78. Zagal JH, Páez M, Silva JF (2006) Fundamental aspects on the catalytic activity of metallomacrocyclics for the electrochemical reduction of O<sub>2</sub>. In: Zagal JH, Bedioui F, Dodelet JP (eds) N<sub>4</sub>-macrocyclic metal complexes. Springer, New York, pp 41–82

79. Yeager EB (1984) Electrocatalysts for O<sub>2</sub> reduction. *Electrochim Acta* 29:1527–1537
80. Ohya T, Kobayashi N, Sato M (1987) Bonding character in tetracoordinate (phthalocyanato) iron(II) complexes with electron-withdrawing substituents as studied by Mössbauer spectroscopy. *Inorg Chem* 26:2506–2509
81. Ohya T, Sato M (1997) Mössbauer and visible spectra of iron(III) complexes of para-substituted tetraphenylporphyrins - electronic effects of substituents and axial ligands. *J Inorg Biochem* 67:126, Abstract:126
82. Schlosser K, Hoyer E, Arnold D (1974) Mössbauer Untersuchungen zur  $\pi$ -Rückbindung in Tris(1,2-diimin)eisen (II)chelaten. *Spectrochim Acta* 30:1431–1436
83. Taube R (1974) New aspects of the chemistry of transition metal phthalocyanines. *Pure Appl Chem* 38:427–438
84. Walker FA (2003) Pulsed EPR and NMR spectroscopy of paramagnetic iron porphyrinates and related iron macrocycles: how to understand patterns of spin delocalization and recognize macrocycle radicals. *Inorg Chem* 42:4526–4544
85. Steiger B, Anson FC (1995) Evidence for the importance of back-bonding in determining the behavior of ruthenated cyanophenyl cobalt porphyrins as electrocatalysts for the reduction of dioxygen. *Inorg Chem* 34:3355–3357
86. Zagal JH, Gulppi M, Isaacs M, Cárdenas-Jirón G, Aguirre MJ (1998) Linear versus volcano correlations between electrocatalytic activity and redox and electronic properties of metallophthalocyanines. *Electrochim Acta* 44:1349–1357
87. Alt H, Binder H, Lindner W, Sandstede G (1971) Metal chelates as electrocatalysts for oxygen reduction in acid electrolytes. *J Electroanal Chem* 31:A19–A22
88. Vasudevan P, Santosh MN, Tyagi S (1990) Transition metal complexes of porphyrins and phthalocyanines as electrocatalysts for dioxygen reduction. *Transition Metal Chem Lond* 15:81–90
89. Kozawa A, Zilionis VE, Brodd RJ (1970) Oxygen and hydrogen peroxide reduction at a ferric phthalocyanine-catalyzed graphite electrode. *J Electrochem Soc* 117:1470–1474
90. Okunola A, Kowalewska B, Bron M, Kulesza PJ, Schuhmann W (2009) Electrocatalytic reduction of oxygen at electropolymerized films of metalloporphyrins deposited onto multiwalled carbon nanotubes. *Electrochim Acta* 54:1954–1960
91. Okunola AO, Nagaiah TC, Chen X, Eckhard K, Schuhmann W, Bron M (2009) Visualization of local electrocatalytic activity of metalloporphyrins towards oxygen reduction by means of redox competition scanning electrochemical microscopy (RC-SECM). *Electrochim Acta* 54:4971–4978
92. Collman JP, Denisevich P, Konai Y, Marrocco M, Koval C, Anson FC (1980) Electrode catalysis of the four-electron reduction of oxygen to water by dicobalt face-to-face porphyrins. *J Am Chem Soc* 102:6027–6036
93. Biloul A, Coowar F, Contamin O, Scarbeck G, Savy M, van den Ham D, Riga J, Verbist JJ (1993) Oxygen reduction in an acid medium: Electrocatalysis by CoNpC(1,2)-bilayer impregnated on a carbon black support: effect of loading and heat treatment. *J Electroanal Chem* 350:189–204
94. Collman JP, Wagenknecht PS, Hutchison JE (1994) Cofaciale Bis(metallo)diporphyrine als potentielle molekulare Katalysatoren für Mehrelektronenreduktionen und -oxidationen kleiner Moleküle. *Angew Chem* 106:1620–1639
95. Liu HY, Weaver MJ, Wang C-B, Chang CK (1983) Dependence of electrocatalysis for oxygen reduction by adsorbed dicobalt cofacial porphyrins upon catalyst structure. *J Electroanal Chem* 145:439–447
96. Biloul A, Contamin O, Scarbeck G, Savy M, Palys B, Riga J, Verbist JJ (1994) Oxygen reduction in acid media: influence of the activity of conpc(1,2)-bilayer deposits in relation to their attachment to the carbon black support and role of surface groups as a function of heat treatment. *J Electroanal Chem* 365:239–246
97. Tributsch H, Koslowski U, Dorbandt I (2008) Experimental and theoretical modeling of Fe-, Co-, Cu-, Mn based electrocatalysts oxygen reduction. *Electrochim Acta* 53:2198–2209
98. Wiesener K, Ohms D, Neumann V, Franke R (1989) N<sub>4</sub> Macrocycles as electrocatalysts for the cathodic reduction of oxygen. *Mater Chem Phys* 22:457–475
99. Blomquist J, Helgeson U, Moberg LC, Johansson LY, Larsson R (1982) Electrochemical and Mössbauer investigations of polymeric iron phthalocyanine oxygen electrodes. *Electrochim Acta* 27:1445–1451
100. Scherson DA, Fierro C, Tryk D, Gupta SL, Yeager EB, Eldridge J, Hoffman RW (1985) In situ Mössbauer spectroscopy and electrochemical studies of the thermal stability of iron phthalocyanine dispersed in high surface area carbon. *J Electroanal Chem* 184:419–426
101. Scherson DA, Fierro C, Yeager EB, Kordesch ME, Eldridge J, Hoffman RW, Barnes A (1984) In situ Mössbauer spectroscopy on an operating fuel cell. *J Electroanal Chem* 169:287–302
102. Wiesener K (1986) N<sub>4</sub> chelates as electrocatalysts for the cathodic oxygen reduction. *Electrochim Acta* 31:1073–1078
103. Tanaka AA, Fierro C, Scherson DA, Yeager EB (1989) Oxygen reduction on adsorbed iron tetrapyrrolineporphyrins. *Mater Chem Phys* 22:431–456
104. Tarasevich MR, Radyushkina KA (1989) Pyropolymers of N<sub>4</sub>-complexes: structure and electrocatalytic properties. *Mater Chem Phys* 22:477–502
105. Franke R, Ohms D, Wiesener K (1989) Investigation of the influence of thermal treatment on the properties of carbon materials modified by N<sub>4</sub>-chelates for the reduction of oxygen in acidic media. *J Electroanal Chem* 260:63–73
106. Joyner RW, van Veen JAR, Sachtler WMH (1982) Extended X-ray absorption fine structure (EXAFS) study of cobalt-porphyrin catalysts supported on active carbon. *J Chem Soc Faraday Trans* 78:1021–1028
107. Saha MS, Li R, Sun X, Ye S (2009) 3-D composite electrodes for high performance PEM fuel cells composed of Pt supported on nitrogen-doped carbon nanotubes grown on carbon paper. *Electrochem Commun* 11:438–441
108. Biniak S, Szymanski G, Siedlewski J, Swiatkowski A (1997) The characterization of activated carbons with oxygen and nitrogen surface groups. *Carbon* 35:1799–1810

109. Lacerda M, Lejeune M, Jones BJ, Barklie RC, Bouzerar R, Zellama K, Conway NMJ, Godet C (2002) Electronic properties of amorphous carbon nitride  $a\text{-C}_{1-x}\text{N}_x\text{:H}$  films investigated using vibrational and ESR characterisations. *J NonCryst Solids* 299–302:907–911
110. Chen Y, Wang J, Liu H, Li R, Sun X, Ye S, Knights S (2009) Enhanced stability of Pt electrocatalysts by nitrogen doping in CNTs for PEM fuel cells. *Electrochem Commun* 11:2071–2076
111. Strelko VV, Kuts VS, Thrower PA (2000) On the mechanism of possible influence of heteroatoms of nitrogen, boron and phosphorus in a carbon matrix on the catalytic activity of carbons in electron transfer reactions. *Carbon* 38:1499–1503
112. Kramm UI (2009) Strukturelle Untersuchungen alternativer Katalysatoren für die PEM-BZ – Eine Fe-57-Mößbauer-spektroskopische Studie pyrolysierter Eisenporphyrin-Elektrokatalysatoren. Saarbrücken, Südwestdeutscher Verlag für Hochschulschriften
113. Biddinger EJ, Ozkan US (2007) Methanol tolerance of  $\text{CN}_x$  oxygen reduction catalysts. *Top Catal* 46:339–348
114. Kothandaraman R, Nallathambi V, Artyushkova K, Barton SC (2009) Non-precious oxygen reduction catalysts prepared by high-pressure pyrolysis for low-temperature fuel cells. *Appl Catal B* 92:209–216
115. Liu G, Li X, Ganesan P, Popov BN (2010) Studies of oxygen reduction reaction active sites and stability of nitrogen-modified carbon composite catalysts for PEM fuel cells. *Electrochim Acta* 55:2853–2858
116. Maldonado S, Stevenson KJ (2005) Influence of nitrogen doping on oxygen reduction electrocatalysis at carbon nanofiber. *J Phys Chem B* 109:4707–4716
117. Matter PH, Biddinger EJ, Ozkan US (2007) Non-precious metal oxygen reduction catalysts for PEM fuel cells. *Catalysis* 30:338–366
118. Matter PH, Wang E, Millet J-M, Ozkan US (2007) Characterization of the iron phase in  $\text{CN}_x$ -based oxygen reduction reaction catalysts. *J Phys Chem C* 111:1444–1450
119. Nallathambi V, Li X, Lee J-W, Popov BN (2008) Development of nitrogen-modified carbon-based catalysts for oxygen reduction in PEM fuel cells. *ECS Trans* 16:405–417
120. Sidik RA, Anderson AB, Subramanian NP, Kumaraguru SP, Popov BN (2006)  $\text{O}_2$  reduction on graphite and nitrogen-doped graphite: experimental and theory. *J Phys Chem B* 110:1787–1793
121. Fischer A, Antonietti M, Thomas A (2007) Growth confined by the nitrogen source: synthesis of pure metal nitride nanoparticles in mesoporous graphitic carbon nitride. *Adv Mater Weinheim* 19:264–267
122. Lyth SM, Nabae Y, Moriya S, Kuroki S, Kakimoto M-a, Ozaki J-i, Miyata S (2009) Carbon nitride as a nonprecious catalyst for electrochemical oxygen reduction. *J Phys Chem C* 113:20148–20151
123. Lyth SM, Nabae Y, Islam NM, Kuroki S, Kakimoto M-a, Miyata S (2011) Electrochemical oxygen reduction activity of carbon nitride supported on carbon black. *J Electrochem Soc* 158: B194–B201
124. Nabae Y, Moriya S, Matsubayashi K, Lyth SM, Malon M, Wu L, Islam NM, Koshigoe Y, Kuroki S, Kakimoto M-a, Ozaki K, Miyata S, Ozaki J-i (2010) The role of Fe species in the pyrolysis of Fe phthalocyanine and phenolic resin for preparation of carbon-based cathode catalysts. *Carbon* 48:2613–2624
125. Plejewski P, Fiechter S (2011) Investigation of the oxygen reduction ability of different metal-free C/N-based materials, in preparation
126. Lalonde G, Côté R, Guay D, Dodelet J-P, Weng LT, Bertrand P (1997) Is nitrogen important in the formulation of Fe-based catalysts for oxygen reduction in solid polymer fuel cells? *Electrochim Acta* 42:1379–1388
127. Martinaiou I, Bogdanoff P, Fiechter S, Kramm UI (2011) Synthesis of nitrogen-rich Fe-N-C catalysts for the oxygen reduction in acidic media, in preparation
128. Scherson DA, Tanaka AA, Gupta GP, Tryk DA, Fierro C, Holze R, Yeager EB, Lattimer RP (1986) Transition metal macrocycles supported on high area carbon: pyrolysis-mass spectroscopy studies. *Electrochim Acta* 31:1247–1258
129. Matter PH, Zhang L, Ozkan US (2006) The role of nanostructure in nitrogen-containing carbon catalysts for the oxygen reduction reaction. *J Catal* 239:83–96
130. Bogdanoff P, Herrmann I, Hilgendorff M, Dorbandt I, Fiechter S, Tributsch H (2004) Probing structural effects of pyrolysed CoTMPP-based electrocatalysts for oxygen reduction via new preparation strategies. *J New Mat Elect Syst* 7:85–92
131. Herrmann I, Bogdanoff P, Schmithals G, Fiechter S (2006) Influence of the molecular and mesoscopic structure on the electrocatalytic activity of pyrolysed CoTMPP in the oxygen reduction. *ECS Trans* 3:211–219
132. Herrmann I, Kramm UI, Fiechter S, Bogdanoff P (2009) Oxalate supported pyrolysis of CoTMPP as electrocatalysts for the oxygen reduction reaction. *Electrochim Acta* 54:4275–4287
133. Herrmann I, Kramm UI, Radnik J, Bogdanoff P, Fiechter S (2009) Influence of sulphur on the pyrolysis of CoTMPP as electrocatalyst for the oxygen reduction reaction. *J Electrochem Soc* 156:B1283–B1292
134. Herrmann I, Barkschat C, Fiechter S, Bogdanoff P, Iwata N, Takahashi H (2007) Electrode catalyst for fuel cells, a method of preparing an electrode catalyst for fuel cells, and a polymer electrolyte fuel cell. Patent application number PCT/JP2007/074369
135. Kurak KA, Anderson AB (2009) Nitrogen-treated graphite and oxygen electroreduction on pyridinic edge sites. *J Phys Chem C* 113:6730–6734
136. Kramm UI, Abs-Wurmbach I, Herrmann-Geppert I, Radnik J, Fiechter S, Bogdanoff P (2011) Influence of the electron-density of  $\text{FeN}_4$ -centers towards the catalytic activity of pyrolysed FeTMPPCl-based ORR-electrocatalysts. *J Electrochem Soc* 158:B69–B78
137. Biloul A, Coowar F, Contamin O, Scarbeck G, Savy M, Van den Ham D, Riga J, Verbist JJ (1990) Oxygen reduction in acid media on supported iron naphthalocyanine – effect of isomer configuration and pyrolysis. *J Electroanal Chem* 289:189–201

138. Bouwkamp-Wijnoltz AL, Visscher W, van Veen JAR (1998) The selectivity of oxygen reduction by pyrolysed iron porphyrin supported on carbon. *Electrochim Acta* 43:3141–3152
139. Faubert G, Côté R, Guay D, Dodelet J-P, Dénès G, Bertrand P (1998) Iron catalysts prepared by high-temperature pyrolysis of tetraphenylporphyrins in polymer electrolyte fuel cells. *Electrochim Acta* 43:341–353
140. Faubert G, Lalande G, Côté R, Guay D, Dodelet J-P, Weng LT, Bertrand P, Dénès G (1996) Heat-treated iron and cobalt tetraphenylporphyrins adsorbed on carbon black: physical characterization and catalytic properties of these materials for the reduction of oxygen in polymer electrolyte fuel cells. *Electrochim Acta* 41:1689–1701
141. Gojkovic SL, Gupta S, Savinell RF (1998) Heat-treated iron(III) tetramethoxyphenyl porphyrin chloride supported on high area carbon as an electrocatalyst for oxygen reduction part I: characterization of the electrocatalyst. *J Electrochem Soc* 145:3493–3499
142. Gojkovic SL, Gupta S, Savinell RF (1999) Heat-treated iron(III) tetramethoxyphenyl porphyrin chloride supported on high area carbon as an electrocatalyst for oxygen reduction part II: kinetics of oxygen reduction. *J Electroanal Chem* 462:63–72
143. Gojkovic SL, Gupta S, Savinell RF (1999) Heat-treated iron(III) tetramethoxyphenyl porphyrin chloride supported on high area carbon as an electrocatalyst for oxygen reduction part III: detection of hydrogen-peroxide during oxygen reduction. *Electrochim Acta* 45:889–897
144. Ikeda O, Fukuda H, Tamura H (1986) The effect of heat-treatment on group VIII porphyrins as electrocatalysts in the cathodic reduction of oxygen. *J Chem Soc Faraday Trans* 82:1561–1573
145. Lalande G, Faubert G, Côté R, Guay D, Dodelet J-P, Weng LT, Bertrand P (1996) Catalytic activity and stability of heat-treated iron phthalocyanines for the electroreduction of oxygen in polymer electrolyte fuel cells. *J Power Sources* 61:227–237
146. Savy M, Coowar F, Riga J, Verbist JJ, Bronoël G, Besse S (1990) Investigation of O<sub>2</sub> reduction in alkaline media on macrocyclic chelates impregnated on different supports: influence of the heat treatment on stability and activity. *J Appl Electrochem* 20:260–268
147. Sawaguchi T, Itabashi T, Matsue T, Uchida I (1990) Electrochemical reduction of oxygen by metalloporphyrin ion-complexes with heat-treatment. *J Electroanal Chem* 279:219–230
148. Schulenburg H, Stankov S, Schünemann V, Radnik J, Dorbandt I, Fiechter S, Bogdanoff P, Tributsch H (2003) Catalysts for the oxygen reduction from heat-treated iron(III) tetramethoxyphenylporphyrin chloride: structure and stability of active sites. *J Phys Chem B* 107:9034–9041
149. Sheng T-C, Rebenstorf B, Widelöv A, Larsson R (1992) Pyrolysis of metalloporphyrins part 1: Fourier-transform infrared study of Fe-tetraphenylporphyrin chloride. *J Chem Soc Faraday Trans* 88:477–482
150. Sun G-Q, Wang J-T, Gupta S, Savinell RF (2001) Iron(III) tetramethoxyphenylporphyrin (FeTMPP-Cl) as electrocatalyst for oxygen reduction in direct methanol fuel cells. *J Appl Electrochem* 31:1025–1031
151. Sun G-Q, Wang J-T, Savinell RF (1998) Iron(III) tetramethoxyphenylporphyrin (FeTMPP) as methanol tolerant electrocatalyst for oxygen reduction in direct methanol fuel cells. *J Appl Electrochem* 28:1087–1093
152. Wang H, Côté R, Faubert G, Guay D, Dodelet J-P (1999) Effect of the pre-treatment of carbon black supports on the activity of Fe-based electrocatalysts for the reduction of oxygen. *J Phys Chem B* 103:2042–2049
153. Widelöv A (1993) Pyrolysis of iron and cobalt porphyrins sublimated onto the surface of carbon black as a method to prepare catalysts for O<sub>2</sub> reduction. *Electrochim Acta* 38:2493–2502
154. Lalande G, Côté R, Tamizhmani GD, Dodelet J-P, Dignard-Bailey L, Weng LT, Bertrand P (1995) Physical, chemical and electrochemical characterization of heat-treated tetracarboxylic cobalt phthalocyanine adsorbed on carbon black as electrocatalyst for oxygen reduction in polymer electrolyte fuel cells. *Electrochim Acta* 40:2635–2646
155. Herrmann I, Brüser V, Fiechter S, Kersten H, Bogdanoff P (2005) Electrocatalysts for oxygen reduction prepared by plasma treatment of carbon-supported cobalt tetramethoxyphenylporphyrin. *J Electrochem Soc* 152:A2179–A2185
156. Herrmann I, Kramm UI, Fiechter S, Brüser V, Kersten H, Bogdanoff P (2010) Comparative study of the carbonisation of CoTMPP by low temperature plasma and by heat-treatment. *Plasma Process Polym* 7:515–526
157. Harnisch F, Savastenko NA, Zhao F, Steffen H, Brüser V, Schröder U (2009) Comparative study on the performance of pyrolyzed and plasma-treated iron(III) phthalocyanine-based catalysts for oxygen reduction in pH neutral electrolyte solutions. *J Power Sources* 193:86–92
158. Atanassov P (2005) Non-platinum electrocatalysts for fuel cells. *FC Techn, Aiche*
159. Garsuch A, d'Eon R, Dahn T, Klepel O, Garsuch RR, Dahn JR (2008) Oxygen reduction behaviour of highly porous non-noble metal catalysts prepared by a template-assisted synthesis. *J Electrochem Soc* 155:B236–B243
160. Jaouen F, Herranz J, Lefèvre M, Dodelet J-P, Kramm UI, Herrmann I, Bogdanoff P, Maruyama J, Nagaoka T, Garsuch A, Dahn JR, Olson TS, Pylypenko S, Atanassov P, Ustinov EA (2009) A cross-laboratory experimental review of non-noble-metal catalysts for oxygen electro-reduction. *Appl Mater Interfaces* 1:1623–1639
161. Olson TS, Chapman K, Atanassov P (2008) Non-platinum cathode catalyst layer composition for single membrane electrode assembly proton exchange membrane fuel cell. *J Power Sources* 183:557–563
162. Koslowski UI, Herrmann I, Bogdanoff P, Barkschat C, Fiechter S, Iwata N, Takahashi H, Nishikoro H (2008) Evaluation and analysis of PEM-FC performance using non-platinum cathode catalysts based on pyrolysed Fe- and Co-porphyrins – influence of a secondary heat-treatment. *ECS Trans* 13:125–141

163. Kramm UI, Herrmann I, Fiechter S, Zehl G, Zizak I, Abs-Wurmbach I, Radnik J, Dorbandt I, Bogdanoff P (2009) On the influence of sulphur on the pyrolysis process of FeTMPP-Cl-based electro-catalysts with respect to oxygen reduction reaction (ORR) in acidic media. *ECS Trans* 25:659–670
164. Maldonado S, Stevenson KJ (2004) Direct preparation of carbon nanofiber electrodes via pyrolysis of iron(II) phthalocyanine: electrocatalytic aspects for oxygen reduction. *J Phys Chem B* 108:11375–11383
165. Lefèvre M, Proietti E, Jaouen F, Dodelet J-P (2009) Iron-based catalysts with improved oxygen reduction activity in polymer electrolyte fuel cells. *Science* 324:71–74
166. Wood TE, Tan Z, Schmoeckel AK, O' Neill D, Atanasoski R (2008) Non-precious metal oxygen reduction catalyst for PEM fuel cells based on nitroaniline precursor. *J Power Sources* 178:510–516
167. Kramm UI, Herrmann-Geppert I, Bogdanoff P, Abs-Wurmbach I, Fiechter S (2011) Effect of an ammonia treatment on the structural composition and ORR activity of Fe-N-C catalysts. *J Phys Chem C*, in preparation
168. Charreteur F, Jaouen F, Ruggeri S, Dodelet J-P (2008) Fe/N/C Non-precious catalysts for PEM fuel cells: influence of the structural parameters of pristine commercial carbon blacks on their activity for oxygen reduction. *Electrochim Acta* 53:2925–2938
169. Charreteur F, Ruggeri S, Jaouen F, Dodelet J-P (2008) Increasing the activity of Fe/N/C catalysts in PEM fuel cell cathodes using carbon blacks with a high disordered carbon content. *Electrochim Acta* 53:6881–6889
170. Dodelet J-P (2006) Oxygen reduction in PEM fuel cell conditions: heat-treated non-precious metal-N<sub>4</sub> macrocycles and beyond. In: Zagal JH, Bedioui F, Dodelet JP (eds) N<sub>4</sub>-macrocyclic metal complexes: electrocatalysis, electrophotocatalysis & biomimetic electroanalysis. Springer, New York, pp 83–147
171. Herranz J, Lefèvre M, Larouche N, Stansfield B, Dodelet J-P (2007) Step-by-step synthesis of non-noble metal electrocatalysts for O<sub>2</sub> reduction under proton exchange membrane fuel cell conditions. *J Phys Chem C* 111:19033–19042
172. Jaouen F, Dodelet J-P (2007) Turn-over frequency of O<sub>2</sub> electro-reduction for Fe/N/C and Co/N/C catalysts in PEFCs. *Electrochim Acta* 52:5975–5984
173. Matter PH, Wang E, Arias M, Biddinger EJ, Ozkan US (2006) Oxygen reduction reaction catalysts prepared from acetonitrile pyrolysis over alumina-supported metal particles. *J Phys Chem B* 110:18374–18384
174. Bron M, Radnik J, Fieber-Erdmann M, Bogdanoff P, Fiechter S (2002) EXAFS, XPS and electrochemical studies on oxygen reduction catalysts obtained by heat treatment of iron phenanthroline complexes supported on high surface area carbon black. *J Electroanal Chem* 535:113–119
175. Nallathambi V, Lee J-W, Kumaraguru SP, Wu G, Popov BN (2008) Development of high performance carbon composite catalyst for oxygen reduction reaction in PEM proton exchange membrane fuel cells. *J Power Sources* 183:34–42
176. Wu T-M, Lin S-H (2006) Characterization and electrical properties of polypyrrole/multiwalled carbon nanotube composites synthesized by in situ chemical oxidative polymerization. *J Polym Sci Pol Phys* 44:1413–1418
177. Bashyam R, Zelenay P (2006) A class of non-precious metal composite catalysts for fuel cells. *Nat Lond* 443:63–66
178. Bron M, Fiechter S, Hilgendorf M, Bogdanoff P (2002) Catalysts for oxygen reduction from heat-treated carbon-supported iron phenanthroline complexes. *J Appl Electrochem* 32:211–216
179. Côté R, Lalonde G, Guay D, Dodelet J-P (1998) Influence of nitrogen-containing precursors on the electrocatalytic activity of heat-treated Fe(OH)<sub>2</sub> on carbon black for O<sub>2</sub> reduction. *J Electrochem Soc* 145:2411–2418
180. Jaouen F, Lefèvre M, Dodelet J-P, Cai M (2006) Heat-treated Fe/N/C Catalysts for O<sub>2</sub> electroreduction: are active sites hosted in micropores? *J Phys Chem B* 110:5553–5558
181. Jaouen F, Marcotte S, Dodelet J-P, Lindbergh G (2003) Oxygen reduction catalysts for polymer electrolyte fuel cells from the pyrolysis of iron acetate adsorbed on various carbon supports. *J Phys Chem B* 107:1376–1386
182. Lefèvre M, Jaouen F, Dodelet J-P, Li XH, Chen K, Hay A (2006) Fe-based catalyst for oxygen reduction: functionalization of carbon black and importance of the microporosity. *ECS Trans* 3:201–210
183. Maruyama J, K-i S, Kawaguchi M, Abe I (2004) Influence of activated carbon pore structure on oxygen reduction at catalyst layers supported on rotating disk electrodes. *Carbon* 42:3115–3121
184. Proietti E, Ruggeri S, Dodelet J-P (2008) Fe-based electrocatalysts for oxygen reduction in PEMFCs using ball-milled graphite powder as a carbon support. *J Electrochem Soc* 155:B340–B348
185. Wang P, Ma Z, Zhao Z, Jia L (2007) Oxygen reduction on the electrocatalysts based on pyrolysed non-noble metal/poly-phenylenediamine/carbon black composites: new insight into the active sites. *J Electroanal Chem* 611:87–95
186. Wu G, Chen Z, Artyushkova K, Garzona FH, Zelenay P (2008) Polyaniline-derived non-precious catalyst for the polymer electrolyte fuel cell cathode. *ECS Trans* 16:159–170
187. He P, Lefèvre M, Faubert G, Dodelet J-P (1999) Oxygen reduction catalysts for polymer electrolyte fuel cells from the pyrolysis of various transition metal acetates adsorbed on 3,4,9,10 perylenetetracarboxylic dianhydride. *J New Mater Electrochem Sys* 2:243–251
188. Bron M, Fiechter S, Bogdanoff P, Tributsch H (2002) Thermogravimetry/mass spectrometry investigations on the formation of oxygen reduction catalysts for PEM fuel cells on the basis of heat-treated iron phenanthroline complexes. *Fuel Cells* 2:137–142
189. Wu G, Artyushkova K, Ferrandon M, Kropf AJ, Myers D, Zelenay P (2009) Performance durability of polyaniline-derived non-precious cathode catalysts. *ECS Trans* 25:1299–1311
190. Kramm UI, Herranz J, Arruda TM, Larouche N, Lefèvre M, Jaouen F, Bogdanoff P, Fiechter S, Abs-Wurmbach I, Stansfield B, Mukerjee S, Dodelet JP (2011) The role of

- Fe-modifications in the activity of Fe/N/C catalysts for the O<sub>2</sub>-reduction in PEM fuel cells, in preparation
191. Herranz J, Jaouen F, Lefevre M, Kramm UI, Proietti E, Dodelet JP, Bogdanoff P, Fiechter S, Abs-Wurmbach I, Bertrand P, Arruda TM, Mukerjee S (2011) Unveiling N-protonation and anion binding effects on Fe/N/C-catalysts for O<sub>2</sub> reduction in PEM fuel cells. *J Phys Chem C*, submitted (2011-05-10)
  192. Lefèvre M, Proietti E, Jaouen F, Dodelet J-P (2009) Iron-based catalysts for oxygen reduction in PEM fuel cells: expanded study using the pore-filling method. *ECS Trans* 25:105–115
  193. Lee K, Zhang L, Lui H, Hui R, Shi Z, Zhang J (2009) Oxygen reduction reaction (ORR) catalyzed by carbon-supported cobalt polypyrrole (Co-PPy/C) electrocatalysts. *Electrochim Acta* 54:4704–4711
  194. Liu H, Shi Z, Zhang J, Zhang L, Zhang J (2009) Ultrasonic spray pyrolyzed iron-polypyrrole mesoporous spheres for fuel cell oxygen reduction electrocatalysts. *J Mater Chem* 19:468–470
  195. Gasteiger HA, Markovic NM (2009) Just a dream – or future reality? *Science* 324:48–49
  196. Gasteiger HA, Kocha SS, Sompalli B, Wagner FT (2005) Activity benchmarks and requirements for Pt, Pt-alloys, and non-Pt oxygen reduction catalysts for PEMFCs. *Appl Catal B Environ* 56:9–35
  197. Maruyama J, Abe I (2007) Fuel cell cathode catalyst with heme-like structure formed from nitrogen of glycine and iron. *J Electrochem Soc* 154:B297–B304
  198. Nallathambi V, Wu G, Subramanian NP, Kumaraguru SP, Lee J-W, Popov BN (2007) Highly active carbon composite electrocatalysts for PEM fuel cells. *ECS Trans* 11:241–247
  199. Bezerra CWB, Zhang L, Liu H, Lee K, Marques ALB, Marques EP, Wang H, Zhang J (2007) A review of heat-treatment effects on activity and stability of PEM fuel cell catalysts for oxygen reduction reaction. *J Power Sources* 173:891–908
  200. Ekström H, Hanarp P, Gustavsson M, Fridell E, Lundblad A, Lindbergh G (2006) A novel approach for measuring catalytic activity of planar model catalysts in the polymer electrolyte fuel cell environment. *J Electrochem Soc* 153:A724–A730
  201. Liu G, Li X, Ganesan P, Popov BN (2009) Development of non-precious metal oxygen-reduction catalysts for PEM fuel cells based on N-doped ordered porous carbon. *Appl Catal B Environ* 93:156–165
  202. Ma Z-F, Xie X-Y, Ma X-X, Zhang D-Y, Ren Q, Heß-Mohr N, Schmidt VM (2006) Electrochemical characteristics and performance of CoTMPP/BP oxygen reduction electrocatalysts for PEM fuel cell. *Electrochem Commun* 8:389–394
  203. Maruyama J, Abe I (2006) Carbonized hemoglobin functioning as a cathode catalyst for polymer electrolyte fuel cells. *Chem Mater* 18:1303–1311
  204. Maruyama J, Fukui N, Kawaguchi M, Abe I (2008) Application of nitrogen-rich amino acids to active site generation in oxygen reduction catalyst. *J Power Sources* 182:489–495
  205. Papakonstantinou G, Daletou MK, Kotsifa A, Paloukis F, Katerinopoulou K, Ioannides T, Neophytides SG (2009) Non noble metal electrocatalysts for high temperature PEM fuel cells. *ECS Trans* 25:181–189
  206. Yang J, Liu D-J, Kariuki NN, Chen LX (2008) Aligned carbon nanotubes with built-in FeN<sub>4</sub> active sites for electrocatalytic reduction of oxygen. *Chem Commun* 3:329–331
  207. Jaouen F, Proietti E, Lefèvre M, Chenitz R, Dodelet J-P, Wu G, Chung HT, Johnston C, Zelenay P (2011) Recent advances in non-precious metal catalysis for oxygen-reduction reaction in polymer electrolyte fuel cells. *Energy Environ Sci* 4:114–130
  208. Zelenay P (2010) Advanced cathode catalysts, hydrogen program annual merit review and peer evaluation meeting. Project ID: fc005
  209. Kramm UI, Herrmann-Geppert I, Proietti E, Jaouen F, Lefèvre M, Bogdanoff P, Dodelet J-P, Fiechter S (2011) Effect of a high-energy ball milling on the PEM-fuel-cell performance of Fe-N-C catalysts for the oxygen reduction. *J Power Sources*, in preparation
  210. Beck F (1973) Voltammetrische Untersuchungen an elektrokatalytisch wirksamen Phthalocyaninen und Tetraazaannulenen in konzentrierter Schwefelsäure. *Ber Bunsen Ges fuer Phys Chem* 77:353–364
  211. Meier H, Albrecht W, Tschirwitz U, Zimmerhackl E (1973) Zum Einfluß der Leitfähigkeit von Phthalocyaninen bei der Elektrokatalyse in Brennstoffzellen. *Ber Bunsen Ges fuer Phys Chem* 77:843–849
  212. Bae IT, Scherson DA (1998) In situ X-ray absorption of a carbon monoxide-iron porphyrin adduct adsorbed on high area in an aqueous electrolyte. *J Phys Chem B* 102:2519–2522
  213. Appleby AJ, Savy M, Caro P (1980) The role of transition element multiple spin crossover in oxygen transport and electroreduction in porphyrin and phthalocyanine structures. *J Electroanal Chem* 111:91–96
  214. El Hourch A, Belcadi S, Moisy P, Crouigneau P, Léger JM, Lamy C (1992) Electrocatalytic reduction of oxygen at iron phthalocyanine modified polymer electrodes. *J Electroanal Chem* 339:1–12
  215. Lefèvre M, Dodelet J-P (2003) Fe-based catalysts for the reduction of oxygen in polymer electrolyte membrane fuel cell conditions: determination of the amount of peroxide released during electroreduction and its influence on the stability of the catalysts. *Electrochim Acta* 48:2749–2760
  216. Ziegelbauer JM, Gatewood D, Gullá AF, Ramaker DE, Mukerjee S (2006) X-ray absorption spectroscopy studies of water activation on Rh<sub>x</sub>S<sub>y</sub> electrocatalyst for oxygen reduction reaction application. *Electrochem Solid State Lett* 9:A430–A434
  217. Bae IT, Tryk DA, Scherson DA (1998) Effect of heat treatment on the redox properties of iron porphyrins adsorbed on high area carbon in acid electrolytes: An in situ Fe K-Edge X-ray absorption near-edge structure study. *J Phys Chem B* 102:4114–4117
  218. Birry L, Zagal JH, Dodelet J-P (2010) Does CO poison Fe-based catalysts for ORR? *Electrochem Commun* 12:628–631
  219. Borup RL, Meyers J, Pivovar B, Kim YS, Mukundan R, Garland N, Myers D, Wilson M, Garzon F, Wood D, Zelenay P, More K, Stroh K, Zawodzinski T, Boncella J, McGrath JE, Inaba M, Miyatake K, Hori M, Ota K, Ogumi Z, Miyata S, Nishikata A,



- Siroma Z, Uchimoto Y, Yasuda K, K-i K, Iwashita N (2007) Scientific aspects of polymer electrolyte fuel cell durability and degradation. *Chem Rev* 107:3904–3951
220. Young AP, Stumper J, Gyenge E (2009) Characterizing the structural degradation in a PEMFC cathode catalyst layer: carbon corrosion. *J Electrochem Soc* 156:B913–B922
221. Baker R, Wilkinson DP, Zhang J (2008) Electrocatalytic activity and stability of substituted iron phthalocyanines towards oxygen reduction evaluated at different temperatures. *Electrochim Acta* 53:6906–6919
222. Charreteur F, Jaouen F, Dodelet J-P (2009) Iron porphyrin-based cathode catalysts for PEM fuel cells: influence of pyrolysis gas on activity and stability. *Electrochim Acta* 54:6622–6630
223. Meng H, Larouche N, Lefèvre M, Jaouen F, Stansfield B, Dodelet J-P (2010) Iron porphyrin-based cathode catalysts for polymer electrolyte membrane fuel cells: Effect of NH<sub>3</sub> and Ar mixtures as pyrolysis gases on catalytic activity and stability. *Electrochim Acta* 55:6450–6461
224. Wesselmark M, Lagegren C, Lindbergh G (2009) Degradation studies of PEMFC cathodes based on different types of carbon. *ECS Trans* 25:1241–1250
225. Ettingshausen F, Weidner A, Zils S, Wolz A, Suffner J, Michel M, Roth C (2009) Alternative support materials for fuel cell catalysts. *ECS Trans* 25:1883–1892
226. Huang S-Y, Ganesan P, Zhang P, Popov BN (2009) Development of novel metal oxide supported Pt catalysts for Polymer electrolyte membrane and unitized regenerative fuel cells applications. *ECS Trans* 25:1893–1902
227. Shoyama M, Tomimura T, Okubo Y, Nambu H, Lin M-L, Hara K, Fukuoka A, Ishihara A, K-I O (2009) Synthesis of Ta-oxide based nano-sized cathode catalyst on highly ordered mesoporous carbon for PEM fuel cells. *ECS Trans* 25:1903–1908
228. Mukundan R, Borup RL (2009) Visualising liquid water in PEM fuel cells using neutron imaging. *Fuel Cells* 9:499–505
229. Manke I, Hartnig C, Grünerbel M, Kaczerowski J, Lehnert W, Kardjilov N, Hilger A, Banhart J, Treimer W, Strobl M (2007) Quasi-in situ neutron tomography on polymer electrolyte membrane fuel cell stacks. *App Phys Lett* 90:184101-(01–03)

## Poultry Breeding

YOAV EITAN, MORRIS SOLLER

Department of Genetics, The Silberman Life Sciences Institute, The Hebrew University of Jerusalem, Jerusalem, Israel

### Article Outline

Glossary

Definition of the Subject

Introduction: Animal Breeding in Transition

Brief History of Broiler Breeding and Broiler Production

Roadmap of the Entry

Stage 1: Selection for Juvenile Weight for Age (JWfA)

Stage 2: Combined Selection for JWfA and Feed Conversion Ratio (FCR)

Stage 3: Combined Selection for JWfA, FCR, and Proportion of Breast Meat (PBM)

The Role of Breeding and Management in the Improvement in Performance of the Modern Broiler Chicken

Accounting for the Long-Term Continuous Response to Selection and for the Punctuated and Coordinated Appearance of Secondary Effects

Genetic Improvement of Broiler Production Traits and Broiler Welfare

Future Directions

Conclusions and Interpretation

Abbreviations

Bibliography

### Glossary

**Breeding value** The deviation of trait value of an individual from the overall population mean that is due to genetic factors and hence can be transmitted to the progeny.

**Breeding nuclei** The closed and pedigreed flocks at the top of the breeding pyramid within which selection is practiced and from which genetic gains are disseminated to the commercial populations.

**Feed conversion ratio** The weight of food in kg required to produce 1 kg of body weight to market.

**Fitness** The ability of the animal to reproduce and survive.

**Functional traits** Traits that contribute to fitness, e.g., reproductive performance, resistance to pathogens and toxins.

**Genetic variation** Variation in trait expression that is due to variation in structure of the genetic material.

**Heritability** The proportion of total variation in trait expression that is due to genetic variation.

**Index selection** Selection based on combination of estimated breeding value of an individual for a number of production and functional traits weighted according to economic importance and genetic correlations to other traits.

**Mass selection** Selection based on the performance of the individual itself, without consideration of performance of close relatives.

**Metabolic heat and water** Heat and water produced as a by-product of body metabolism.

**Pleiotropic effects** Effects of an individual genetic locus on multiple traits.

**Primary breeders** The commercial organizations that maintain and improve breeding nuclei.

**Production traits** Traits that contribute directly to production of the animal product that reaches the consumer, e.g., JWfA, FCR, and PBM.

**Secondary effect** An unintended effect of selection for one trait on some other trait.

### Definition of the Subject

Animal breeding today stands on the verge of a methodological revolution that may greatly increase the rate of genetic improvement in production traits. This will modify the physiology of the animals and the genetic architecture of the population at an unprecedented rate. What will be the broader consequences of such extreme modification? What pitfalls and dangers may be encountered as this process unfolds? Beginning about 60 years ago, the broiler chicken has been subject to intense and effective selection for juvenile growth rate (60 generations), feed conversion ratio (40 generations), and body composition (20 generations). This entry describes the ways in which the production and functional traits of the individual bird and accompanying management practices have responded to this unremitting selection, with the objective of suggesting what the future might hold for other agricultural animal species as they enter this new stage of intense and highly effective selection for production traits.

### Introduction: Animal Breeding in Transition

Animal breeding today stands in the midst of a methodological revolution based on the availability of microarrays that are capable of genotyping tens and hundreds of thousands of polymorphic markers in a single run, and of low-cost whole genome sequencing that will enable the leading animals in the population to be fully sequenced [1–3]. This has enabled development of deoxyribonucleic acid (DNA)-level diagnostic procedures for highly accurate evaluation of breeding

values for production and functional traits. Since DNA can be obtained readily from all animals from birth (and even before) these procedures allow highly effective and intense selection to be implemented at an early age with very high accuracy, and with equal efficiency in males and females; encompassing sex, age, or phenotype-limited characters such as milk and egg production, longevity, and disease resistance. The result will be a quantum leap in the rate of genetic improvement in production traits with consequent far-reaching modification of body development and physiology.

What will be the effects of such intense and effective selection applied to production traits on overall animal performance and capacities? How will management practices and requirements adapt to these changes? The modern broiler chicken, under selection for production traits for over 60 chicken-generations [4] provides an intimation of what to expect [5–9]. The main broiler production traits such as juvenile growth rate, feed conversion ratio, and body composition come to expression in the juvenile bird, are expressed equally in males and females, and have moderate heritabilities, thus enabling highly effective mass selection based on individual phenotype. Introduction of advanced biometrical methods of selection such as Best Linear Unbiased Predictor (BLUP) and Individual Animal Model has made this phenotype-based selection even more effective. Because of the high fecundity of the female chicken, selection for broiler production traits has been intense, and the response to selection has been enormous, converting the broiler chicken from an expensive luxury to a low-cost major meat supplier for the growing human population worldwide [10]. The historical response of the broiler chicken and its management to this selection, models what can be expected from the application of intense and highly effective selection for production traits to other populations of agricultural animals.

### Brief History of Broiler Breeding and Broiler Production

#### The First Commercial Incubator

Keeping a small number of chickens was always a part of the family farm and rural urban household. These were raised primarily for home egg consumption and minor income from sale of surplus eggs and meat. Until the invention of the first commercial incubator in 1875,

replacement chicks had to be hatched by the hens, so that each farm reared its own replacements. With the availability of commercial incubators, specialized breeding farms and hatcheries came into being to supply high-quality chicks to the family farms. Prior to the development of vent sexing by the Japanese in the 1930s, and of sex-linked feather sexing in the 1960s, it was not possible to distinguish male and female chicks by gender until they were a few weeks of age. Consequently, when rearing chicks for egg production, all chicks had to be reared to the age of a month or more when cockerels could be recognized. Having invested this much in the cockerels, it was economically rewarding to rear them to market weight. Thus, each year with the renewal of layer flocks by chicks hatched in the early spring, cockerels for sale provided an important secondary income stream. At this time, therefore, chicken meat was a rather rare and costly commodity generally available only in the spring coincident with layer flock renewal. However, recognizing the potential for additional economic return, some farms began raising an extra hatch of chicks in the summer months for sale as meat birds.

### ***Salmonella pullorum* Free Flocks, Vitamin D**

Large flocks dedicated to poultry meat production as a main source of farm income do not seem to have become widespread until the 1920s, primarily because of high chick mortality due to *Salmonella pullorum* when large numbers of chicks were reared together. In the early 1920s, a test was developed to identify birds carrying the pathogen, enabling breeders to develop *S. pullorum* free flocks that supplied *S. pullorum*-free chicks. Also about this time (1926), it was found that adding vitamin D to the chick diet in the form of cod liver oil could prevent the development of Rickets in winter broilers. Until then, lack of sunshine had prevented rearing of winter broilers in the Northern states. These two developments made it possible to establish a year round broiler industry [11].

### **Cornish Breed Males, the “Chicken of Tomorrow Contests,” Chick “Sexing”**

Growth of the broiler industry was rather slow, and chicken meat remained an expensive specialty item because of the necessity for hand plucking of the feathers. In 1940 the first commercial feather plucker was invented,

enabling plucking, evisceration, and packing of broiler carcasses on the same production line. This opened the way to development of a dedicated broiler industry. The use of Cornish breed males to cross with dual-purpose females to produce broiler chicks with higher proportion of breast meat (PBM) was introduced in the 1930s, and by 1942 almost all commercial broiler chicks were crossbreds of “Cornish” males to females of one of the dual-purpose breeds (White Rock, New Hampshire or Barred Rock). In the 1940s, some of the leading White Rock breeders began to develop a bird with high juvenile growth rate specifically for broiler production, while the colored stocks were still bred for the dual-purpose market. From 1946 to 1951, the A&P supermarket chain sponsored a series of “Chicken of Tomorrow” contests [12] in which chicks of leading breeders were reared to 12 weeks of age and compared. By the third of these contests it was clear that the White Rock female, specifically bred for broiler production crossed to Cornish type male was the leading broiler chicken, widely preferred to crosses to dual-purpose colored females. In addition to superior growth rate, the White Rock crosses also gave a clean attractive plucked carcass, in contrast to the unsightly dark-colored pinfeathers remaining on the plucked carcass of the dark-feathered birds.

Thus, by 1952 all the components of the modern broiler chicken were in place. In this year for the first time, specially bred broilers surpassed surplus farm cockerels of the layer flocks as the main source of chicken meat in the USA. With the advent of specialized broiler lines, it was no longer economical to rear layer strain cockerels for meat, as the cockerel of the layer lines, even of dual-purpose breeds, could not achieve the economic returns of the specialized broiler lines. Vent and later feather sexing enabled the male chicks of the layer flocks to be identified at hatch, so that there was no economic incentive to rear these chicks. As a direct consequence, the poultry breeding industry was able to separate into the distinct and completely independent layer and broiler components found today, and intensive broiler breeding focused on the traits of importance for low-cost broiler production commenced in earnest.

### **Roadmap of the Entry**

Broiler breeding experienced three stages with respect to the production traits that were the primary selection

objectives. In the first stage, the primary production objective was rapid juvenile growth rate, with some attention to breast conformation. In the second stage, increasing attention was paid to feed conversion ratio in addition to juvenile growth rate [13]. In the third stage, continuing to the present time, increasing the proportion of breast meat in the carcass was added as a third major production goal. At all stages except the very first, good reproductive performance of the female line and fertility of the male line were strong secondary selection objectives. The need to maintain female reproductive performance led in the 1970s to the introduction of crosses between independent White Rock female lines to provide a heterotic lift to performance. Thus, the typical commercial broiler chick today is the product of a three-way cross: A pair of purebred “female lines, of dual-purpose White Rock origin,” is crossed to produce the female parent, and this in turn is crossed to males of a purebred “male line of Cornish breed origin” to produce the commercial chick. With the passage of time, selection for health and welfare traits that were negatively affected by the intense selection for production traits has become an increasingly important component of the broiler selection programs.

Each of these three major stages in genetic development of the modern broiler will be considered in turn, describing first the achievements in terms of the primary production goals, followed by a description of the associated secondary genetic effects on reproductive performance (hatchability and egg numbers) and health (with emphasis on leg and metabolic disorders), and of the management modifications needed to accommodate both the primary genetic effects and their secondary consequences. A common feature of the secondary consequences was inability to predict them in prospect, but ability to explain their appearance in retrospect as a plausible consequence of the changes in primary traits. This will be followed by an overall evaluation of the relative contributions of the breeder and of management to performance of the modern broiler, and a review of possible sources of the genetic variation that has enabled the continuous long-term response to selection for the three major production traits. Welfare issues raised by the genetic and management changes will be addressed, and a framework presented for interpreting the

accompanying genetic and management changes in terms of a “resource allocation” model.

### **Stage 1: Selection for Juvenile Weight for Age (JWfA)**

#### **Importance of Juvenile Weight for Age**

A day-old chick of one of the dual-purpose breeds that served as the founders for broiler development, typically required up to 90–120 days to reach market weight [4]. This long grow-out period resulted in high feed, labor, veterinary, and capital costs per chick and increased losses by mishap or disease. The number of days to reach market weight is primarily determined by juvenile growth rate – birds growing more rapidly reach market weight earlier. Consequently, increasing juvenile growth rate was the first objective in the development of the modern broiler. The heaviest birds at any given age are obviously the birds that would have reached target weight earliest, so that selection for “age to reach a given target weight” (age for weight) could be achieved by selection for the much more conveniently measured “weight at a given age” (weight for age).

#### **Rapid and Powerful Response to Selection for JWfA**

Juvenile weight for age (JWfA) turned out to be an ideal target for selection. The trait had high heritability, meaning that much of the variation in the trait was due to genetic factors. It was easily measured by simply weighing the birds at a given age. It came to expression at an early age equally in males and females. Due to the very high fecundity of hens, only a small proportion of the female chicks and an even smaller proportion of the male chicks produced by a hen had to be retained to renew the breeding population. Consequently, selection was very intense. The response to selection for JWfA was magnificent and continues to the present time yielding a decrease of 1 or 2 days in age for market weight per generation of selection [14] so that the modern broiler reaches market weight at about 35–40 days. This tremendous reduction in market age has reduced broiler-rearing costs in a corresponding manner, bringing chicken from a high-cost luxury food to the lowest-cost meat producer on the market [10]. It should be noted that even during this early period,

attention was also paid to breast conformation of the broiler chick [6]. This was needed in order to produce a more attractive whole carcass with a less prominent keel bone on the supermarket shelf. Usually this was done by “touch,” i.e., simply handling the birds already selected on JWfA and rejecting those with the sharpest keel bone.

### Secondary Effects of Selection for JWfA (I): Excess Adiposity

The first secondary effect of the selection for JWfA, a marked increase in the adiposity of the commercial broiler at slaughter, was noted by the mid-1960s [6, 15]. This was unexpected, since deposition of a gram of fat is more than twice as costly in caloric terms as deposition of a gram of muscle. Thus, the heaviest birds would be expected to be those that had deposited the least fat. Consequently, selection for JWfA should have reduced the proportion of food intake devoted to fat stores, and in this way reduced the proportion of fat content of the carcass. The situation turned out to be more complex. Depositing body mass in a growing bird requires an excess of food intake over body maintenance requirements. Given unlimited access to feed, the magnitude of this excess is a function of appetite. Thus, JWfA is primarily a function of appetite and gross food intake. The birds with the highest intake put on the most weight. However, this high intake introduces a second factor into the picture. Any excess of intake over maintenance in the juvenile is partitioned between lean body growth (muscle and internal organs) and fat stores. It turns out that the coefficient of partition is itself a function of the excess. When excess is small, almost all goes to build lean body mass (low fat:lean coefficient) but as the excess increases, a greater proportion goes to fat stores (higher fat:lean coefficient). Thus, the animals presenting highest JWfA are primarily those with the highest food intake but these also have the highest fat:lean coefficient, and hence show increased adiposity. The increase in carcass fat content made the carcass unsightly and unattractive to consumers. In addition, the abdominal fat-pad that was discarded in the slaughterhouse represented a waste of feed and was clearly reducing feed conversion ratio on a salable meat basis. This led in the early 1970s to various attempts to select directly against excess fat, e.g., by measuring skin

thickness. These approaches were not successful, but control of fat content eventually came from an unexpected direction, as will be seen later.

### Secondary Effect of Selection for JWfA (II): EODES in Broiler Breeder Females

A second secondary effect of the intense selection for JWfA was not long in coming and presented as the first of a series of metabolic disorders (see Box 1 for definition and explication) that have afflicted the broiler industry. The disorder manifested as a marked reduction in reproductive performance of the female broiler parent. Egg production dropped precipitously, from over 200 hatching eggs per hen to 120 and below, accompanied by a reduction in fertility and in the hatchability of fertilized eggs. Egg-laying pattern was erratic with many eggs laid outside of the usual time window, and there was a marked increase in double-yolked or otherwise defective eggs. The condition as a whole was termed “Erratic Oviposition Defective Egg Syndrome, EODES” [17]. An outstanding feature was the presence of multiple hierarchies of large yellow follicles in the ovary. The broiler industry was on the verge of collapse, when it was found that restricting (i.e., limiting) feed intake of the broiler pullet during her growth period could completely prevent the problem [18–20]. With this discovery, feed restriction became an essential part of broiler breeder management, both for males and females. With the passage of generations, it has been necessary to increase the degree of feed restriction from 70% of *ad libitum* initially to about 35% of *ad libitum* today. Since EODES is elicited by *ad libitum* feeding and prevented by restricting feed intake, it is plausible that the proximal cause of this disorder was the increased adiposity of the *ad libitum* fed broiler pullet at entry into lay. Excess adiposity is known to interfere with proper reproductive function in many species. The excess adiposity in turn is due at least in part to disruption of the normal homeostatic control over appetite as a consequence of which *ad libitum* fed female-line broiler pullets continue to eat to excess during the entire pullet rearing stage. The hypothalamic appetite control center may not be fully functional in broilers, since lesions of the hypothalamic appetite control center that lead to hyperphagia in layers, do not increase feed consumption in broilers [21].

### Box 1 Metabolic Disorders

Metabolism refers to all of the myriad chemical reactions taking place in the cells of a living organism that utilize the chemical and energetic content of ingested food to build and maintain normal body structure and function. These cellular chemical reactions are actively coordinated with one another across the body of the organism and also with the external environment through an intricate network of sensors and regulatory feedback loops so as to maintain life (i.e., stable body function). An outstanding feature of the metabolic system is the ability to actively adapt and modify body structure and function so as to maintain life even under challenging environments.

The metabolic system consists of many thousands of structural, functional, and regulatory components composed of protein or ribonucleic acid (RNA) molecules. All of these are constructed according to instructions coded in the hereditary material of the cell, i.e., its DNA. Mutations in DNA can change the structure or function of the components of the metabolic system. When this happens, the metabolic system may not be able to preserve normal function, and development terminates in the death of the embryo, or in birth of an individual defective in one or more aspects of normal structure and function. Metabolic derangements of this sort are termed "Metabolic Diseases," and there are many hundreds of different hereditary diseases that fall into this category.

Normal metabolism may also be disrupted when the diet does not include essential components that cannot be synthesized by the body but are needed to build the elements of the metabolic system, or conversely, when the diet contains various elements or compounds in toxic excess. Metabolic derangements of this sort are termed "Metabolic deficiency or toxicity disorders". Examples of broiler "Deficiency Disorders" are Rickets, resulting from Vitamin D or mineral deficiencies, and Crazy Chick Disease, resulting from Vitamin E deficiency. Examples of broiler "Toxicity Disorders" are Fluoride and Vanadium toxicity caused by phosphate contamination of feeds, or Selenium toxicity caused by feeding cereals grown on Selenium-rich soil.

The Metabolic Syndrome is a novel form of Metabolic Disorder, which has recently become very

prominent in human populations [16]. This is a constellation of pathologies that increase the risk of atherosclerotic vascular disease and Type II (late onset) diabetes. Risk factors for development of Metabolic Syndrome include obesity, psychosocial stress, sedentary lifestyle, and age. There are no identified genetic lesions, dietary deficiencies, or toxins. Thus, all elements of the metabolic system are present and at an individual level appear to be functioning normally. The central role of obesity in the etiology of the syndrome is of special interest. A moderate degree of adiposity is a normal part of body structure. The very high adiposity defining obesity is thus a normal body structure taken to extreme. Apparently, this excess alone unbalances the metabolic system and imposes a regulatory burden to maintain stable function. When the adiposity load is further exacerbated by the regulatory load imposed by psychosocial stress, sedentary life style, and age, the metabolic regulatory system is unable to cope and the pathologies characteristic of the Metabolic Syndrome emerge. This example indicates that whenever some body structure or function is brought to extreme development, this may impose an overload on the metabolic regulatory system resulting in manifestation of Metabolic Regulatory Disorders, with specific pathology depending on the type of overload.

The reduced fertility that accompanied the reduction in egg number apparently resulted in part at least, from impairment of male mating performance, perhaps due to sheer physical difficulty of the obese males and females to handle the mechanics of courting and mating or possibly due to hormonal effects on semen quality of the excess adipose tissue in the male. Feed restriction corrected the reduced fertility as well, so that under pullet and male feed restriction, egg production and fertility returned to high levels.

#### Secondary Effect of Selection for JWfA (III): Increase in Egg Weight and Hatchability Problems

A third secondary consequence of selection for JWfA was an overall increase in egg weight [14] particularly in older hens. The primary reason for this was the increase in mature size of the broiler female, even

under feed restriction, as a direct consequence of selection for JWfA. In addition, there may have been indirect selection for large egg size, as a larger egg delivers a larger chick at hatch and higher final weight at market age. The increase in egg size had important effects on incubator management, deriving from the dual function of the incubator in embryo development. The small early stage embryo generates very little internal metabolic heat and must be warmed by the incubator, while the larger and rapidly growing later-stage embryo generates large amounts of metabolic heat and water, both of which must be removed by the incubator (see [Box 2](#)). Achieving these multiple goals requires careful control of incubator temperature, humidity, and ventilation (THV). Matters are complicated by the fact that even under the best conditions there is variation in the specific THV conditions at different points within the incubator due to variation in airflow over the eggs, and location of the egg with respect to incubator inlets and outlets. There is also variation in egg size and in shell conductance for water and heat. All of these contribute to variation in internal embryo temperature and water content among the population of eggs within the incubator at any given time. As a result, while most eggs will be comfortably within their optimum temperature and water content, others may lie close to outside limits for one or both of these factors.

Larger egg size impacts negatively on all of the multiple incubator functions. At the early embryo stages, the larger egg makes it more difficult to achieve optimal temperature throughout the egg mass. At the later embryo stages, the larger egg accommodates a larger embryo that generates more metabolic heat and water toward the end of embryonic development. At the same time, the surface/volume ratio of the larger egg is reduced making it more difficult for the embryo to dissipate excess metabolic heat and water [22]. Taken together with variability in TMV conditions within the incubator, and in size and shell conductance of the eggs, an increasing proportion of eggs find themselves outside of the optimal parameter range for development and hatching, yielding grade B chicks, and lower overall hatching proportions. This is exacerbated by the increased difference in egg size between young and old breeder hens that increases variability in egg size and shell conductance within the incubator when eggs of

### Box 2 Incubator Management

During the early part of incubation, the embryo generates very little metabolic heat, and must be warmed by the incubator. But during the final week of incubation, the large and rapidly growing embryo generates large amounts of metabolic heat, and must be cooled by the incubator. Both objectives are achieved by setting the incubator air temperature at about 37.5°C. This is sufficient to warm the early embryos, and at the same time, with adequate air movement, will cool the later embryos. Thus, the same incubator (called a “multistage” incubator) was able to accommodate eggs in different stages of development, warming the early embryo eggs and simultaneously cooling the late-embryo eggs. In fact with the multistage incubator, the excess heat produced by the late embryos was used to warm the early embryos.

In addition to producing metabolic heat, the developing embryo also produces water and carbon dioxide. Both carbon dioxide and water escape through the eggshell. However, loss of water must be adjusted carefully. Loss of too much water will dehydrate the chorion-allantoic membrane, interfering with importation of oxygen and exportation of metabolic waste, resulting in death of the embryo. It is equally important that the embryo lose more water than it generates (net loss of about 12% of egg weight from beginning to end of incubation). Otherwise, there will be insufficient air space formed for the chick to take its first breath and the chick itself will remain too large to move around inside the shell when trying to pip. In this case, the chick may be unable to break the membrane between itself and the shell, suffocating or drowning to death. Water loss by the embryo is adjusted primarily by adjusting the humidity of the incubation chamber; 50% humidity is close to optimal.

different-aged flocks are incubated together. Thus, as egg weight increased, incubators had to be equipped with more effective ventilation systems, and conditions had to be monitored more closely than in the past to ensure minimum variation and close adherence to optimal temperature and humidity throughout the incubator interior.

### **Secondary Effect of Selection for JWfA (IV): Skeletal Problems in Commercial Broiler**

Initially, broiler chicks were allowed to free access to feed during daylight hours from the first day post-hatch, with the feed provided in the form of coarsely ground grain (“mash” type feed). However, by the 1970s as growth rate increased, the chicks were unable to consume the amounts of feed needed to meet their growth potential. This was solved by compressing and heating the mash to form pellets, enabling the bird to ingest more feed in a shorter time; and by extending hours of light to 23 h per day to provide more time for feeding. It was later found that equivalent results at lower electricity cost could be obtained by alternating 1–2 h of light with 2–4 h of dark (“intermittent light”) with the added advantage of reducing energy expenditure for locomotor activity by the growing broiler. Obviously implementing a light-control program of this sort required a shift to light-controlled housing.

However, as chick growth rate increased, the skeleton could not develop rapidly enough to accommodate the large chick body, resulting in many skeletal and leg problems [13, 23, 24]. These include: infected hocks (from staphylococcus, coliform, or viral infections); twisted legs presenting as valgus distortion (knock-kneed: hocks in feet out) or varus distortion (bow-legged: hocks out, feet in); and tibial dyschondroplasia, in which failure of normal chondrolysis and ossification leaves the end-bone epiphyseal plates at the femoral–tibial junction prone to fracture, infection, and abnormal development resulting in lameness, and swelling of the femoral–tibial joints. These problems did not appear during rearing of broiler parent flocks from day old to maturation with restricted feeding, indicating that the skeletal and leg problems were a result of the very rapid and unbalanced growth rate of the broiler chick, rather than being innate to the animal. This led to more tailored feeding and lighting programs for young broiler chicks, holding growth below the potential in the early part of the growth period, by returning to mash type feed or otherwise decreasing the nutritional content of the diet, and by decreasing hours of light [23, 25]. In the final weeks of the broiler growth period, after the skeleton and body frame were more solidly formed, photoperiod was increased to maximize feed intake, and diet was shifted

back to nutritionally dense pelleted form to allow full compensatory growth. Skeletal abnormalities were a leading cause of mortality and carcass condemnations in broiler production in the 1970s. At the present time, however, due to genetic selection to reduce their incidence and formulation of suitable management protocols, skeletal deformities are well controlled.

### **Stage 2: Combined Selection for JWfA and Feed Conversion Ratio (FCR)**

Random sample tests of broiler chickens from different breeding firms were instituted in the 1960s to provide unbiased information as to broiler chick quality. In addition to differences in JWfA, these also showed large differences in feed conversion ratio (FCR) among the various breeders, independent to an appreciable degree of juvenile growth rate. This was surprising, since the metabolic efficiency of feed utilization was thought to be highly correlated with evolutionary fitness and hence it was anticipated that there would be little genetic variation in the trait. Because of the economic importance of FCR, starting in the 1970s breeders included this in their breeding programs as a second major objective along with JWfA, with considerable success.

Further consideration led to the realization that although the metabolic efficiency of catabolism and anabolism may not vary, variation in motor activity, carcass adiposity, and possibly protein turnover rate [14, 26, 27] may all contribute to genetic variation in FCR. Consequently, a positive albeit unexpected secondary effect of selection for reduced FCR was a gradual but steady reduction in carcass adiposity [6, 9], apparently by reducing the fat:lean partition coefficient at high intake. Thus, at the present time, excess broiler fat content of the commercial broiler is no longer considered a problem by the broiler industry.

### **Secondary Effect of Continued Selection for JWfA and FCR (V): Adverse Effects on Reproductive Performance of Broiler Female and Male Parents**

The late 1970s and 1980s, however, brought about a number of additional problems, apparently secondary effects of the continued selection for JWfA and FCR. First of all, female and male reproductive performance again became problematic. As noted, to control EODES, female pullets were reared under feed



restriction. Initially, birds of the reproductive flock were shifted to *ad libitum* feeding to induce entry into lay, and remained on *ad libitum* feeding during the entire reproductive period. However, in the 1970s while peak production remained high, post-peak production dropped rapidly. This was apparently due to continued effects of the loss of appetite control that led to overeating and fat accumulation during the laying period, by both females and males. As a consequence, feed restriction was extended into the laying period itself [18, 19]. This continues to the present time and is effective in maintaining egg production.

The extension of feed restriction to the laying period had a further positive effect in that egg weight is closely related to actual body weight and feed consumption, so that egg weight can be closely controlled by monitoring body weight and feed intake during the reproductive period. Control of egg weight in this manner, however, is limited by the ever-increasing optimum mature body weight of the female parent even under feed restriction. This is due in all likelihood to an increase in threshold weight for sexual maturity that has all along accompanied the increase in JWfA.

#### **Secondary Effect of Continued Selection for JWfA and FCR (VI): Increased Water Consumption, Nipple Drinkers**

The increased daily feed consumption of the rapidly growing broiler chick required a proportional increase in water consumption to enable the gizzard and digestive system of the bird to deal effectively with the dry feed. Drinking the additional water from the standard “Bell drinker” resulted in increased water spillage and wet litter causing a host of veterinary and welfare problems, such as foot pad and breast dermatitis, and ammonia production interfering with proper breathing and contributing to pulmonary hypertension syndrome (PHS). This was solved in the early 1980s by the introduction of the “nipple drinker” in which water drops are drawn directly into the chicks’ throat in this way greatly reducing spillage and the attendant problems with wet litter.

#### **Secondary Effect of Continued Selection for JWfA and FCR (VII): Delayed Entry into Lay**

In the 1980s, a new set of problems arose related to entry of the birds into lay. Under natural conditions,

onset of sexual maturity in chickens is controlled by day length (photoperiod). Male and female birds enter sexual maturity in the spring, under the stimulus of gradually increasing photoperiod. Normally, chickens will not enter lay in the fall, due to the negative stimulus of the naturally decreasing photoperiod at this season. In order to have a steady supply of broiler chicks, however, it is necessary to have males and females enter sexual maturity and lay throughout the year. For birds maturing in the fall or winter, this was achieved by extending the photoperiod through the use of supplemental artificial light, added at both ends of the natural day. This produced an artificial spring-type light pattern, and brought the birds into sexual maturity in a reliable manner throughout the year. Beginning in the 1980s, however, onset of sexual maturity was delayed for birds maturing in the fall season of the year, even under the stimulus of supplemental artificial light. Some of the females did not enter lay at all; in others, onset of lay was delayed and peak lay was lower. A management solution was found in the form of “Stimulatory lighting.” For this, the chicks are reared in fully enclosed so-called dark-out houses in which photoperiod is under total artificial control, under a regime of 16 h total darkness and 8 h dim light until it is time for them to enter lay. As noted above, during this period they are also under quantitative feed restriction. To bring the birds into lay (generally, at about 6 months of age), they are moved to the laying pens, and exposed to a photoperiod of 14–16 h of daylight, while at the same time feed quantities are increased rapidly (within a few weeks from 100 g/day to over 160 g/day). The combined stimulus of increased photoperiod and feed quantity brings the birds into sexual maturity and lay.

Initially, dark-out rearing and stimulatory lighting were required only for birds entering lay in the fall. However, with the passage of generations, and continued selection for juvenile growth rate and FCR, problems on entering lay appeared even in birds coming to sexual maturity in the spring of the year under optimal natural lighting. At present, therefore, stimulatory lighting to bring the birds into sexual maturity is required at all seasons of the year [28]. Experimental studies showed that the need for stimulatory lighting was due to reduced innate photosensitivity of the broiler chicks [29], as well as to a direct effect of feed

restriction per se on photosensitivity. Apparently the control systems of the bird interpret feed restriction as indicating that environmental conditions are not yet suitable for chick rearing, and delay onset of lay accordingly.

#### **Secondary Effect of Continued Selection for JWfA and FCR (VIII): Male Overfeeding**

In the mid-1980s, growth and appetite patterns of males and females began to diverge sufficiently, so that even under feed restriction, the males, when fed together with the females, became overweight and lost fertility. This was resolved by the use of special feeders and feed composition for the males, while using barriers preventing the males from having access to the female feeders.

#### **Secondary Effect of Continued Selection for JWfA and FCR (IX): Heat Distress**

Heat is produced by all aspects of body metabolism, including digestion and growth. Producing a given total mass of body weight, therefore, generates a more or less fixed total quantity of metabolic heat to digest the food required to grow this body mass and convert the digested food into body mass. This metabolic heat must be dissipated to avoid symptoms of heat distress. When the grow-out period is long, the daily production of metabolic heat for growth and digestion per unit time is low and readily dissipated by the bird. However, as age to market weight decreases with each generation of selection, this same total metabolic heat is produced in a shorter and shorter grow-out period, so that in each generation the metabolic heat generated per unit time increases [30]. This is particularly true for the final weeks of the grow-out period, when the bulk of body mass is produced.

Generation of metabolic heat interacts strongly with stocking density, a major management determinant of broiler costs through its effect on fixed costs of housing per unit of product. Generally, birds are stocked at a density that gives almost complete ground cover toward the end of the rearing period. Under these conditions, there is maximum radiant heat transmission between birds, and stagnant hot air is trapped between birds and between the birds and the litter. Broilers with slow to moderately rapid growth can be

reared in open sheds, with air circulation controlled by opening and closing window flaps or curtains, as the weather dictates. At optimal ambient temperature of 18–22°C, normal air circulation adjusted in this manner is sufficient to dissipate metabolic heat and the birds are comfortable. As ambient temperature increases, the birds cope by generating evaporative heat loss through panting [31]. For broilers with slow to moderately rapid growth, this is ordinarily sufficient to deal with the stress of normal summer peak heat. When peak heat exceeds ability of the bird to cope by panting, metabolic heat can be reduced by reducing food intake during the hottest part of the day, and by artificially “fogging” or “misting” the birds to increase evaporative heat loss when humidity is less than 70%. However, with the increase in growth rate achieved by the early 1980s, metabolic heat production increased to the point where these palliatives were no longer sufficient and the birds exhibited excessive panting and heat distress under conditions that were previously adequate. Heat distress was particularly acute at high stocking densities toward the end of the growing period, when floor cover by the birds was maximal. This led to reduction in growth rate since the birds reduced feed intake during the hottest part of the day, and in some cases to collapse and death of the bird from hyperthermia. To deal with this, fans were introduced into the chicken sheds to create additional circulation of air inside the sheds and to increase exchange of air between the inside and outside of the sheds. This system worked well, and in combination with misting or fogging when outside temperatures were high, enabled rapid growth to continue even under hot summer weather conditions with good FCR and low mortality.

#### **Secondary Effect of Continued Selection for JWfA and FCR (X): Pulmonary Hypertension Syndrome and Sudden Death Syndrome**

In 1974, a new pathological condition was reported in broiler birds reared at high altitudes. The condition manifested as sudden death, accompanied by “ascites” (accumulation of protein-rich fluid in the body cavity). Initially, death was attributed to the ascites *per se*, and hence the condition was termed “ascites syndrome.” However, further studies showed that the ascites was a secondary consequence of pulmonary hypertension

and the disease is now termed “Pulmonary Hypertension Syndrome” (PHS). With time, PHS manifested in broiler flocks growing at lower altitudes causing high mortality particularly in males of rapidly growing broiler lines from 4 weeks of age. The etiology of the disease has been worked out in detail (Box 3), and it appears to be a direct consequence of continued selection for rapid JWfA, exacerbated by selection for reduced FCR.

At about the same time that PHS manifested, a second metabolic disorder, “Sudden Death Syndrome” (SDS), also became prominent [32, 34, 35]. SDS presents as sudden convulsions, with squawking, violent flapping, and loss of balance. Death ensues within less than a minute from the onset of symptoms, with the bird lying on its back with one or both legs extended. Greatest losses are from 3 to 6 weeks of age, primarily in males. The syndrome is associated with the same factors that induce PHS, namely, high carbohydrate intake, dense housing, very rapid growth, and low feed conversion ratios; but most commonly manifests at an earlier age than PHS. Also similar to PHS, under inducing conditions, SDS can be precipitated by any sudden movements or noises that cause a stress response in the birds. It is plausible therefore, that SDS is an early pathological response to the hypoxia that accompanies rapid growth in broilers (see also Box 3).

From the above, it is clear that incidence of mortality due to PHS and SDS will be increased by any management factor that increases oxygen demands of the animal, such as social stress, low ambient temperature, feeding high-energy food in pellet form that stimulates growth rate; or that limits oxygen supply, such as overcrowding and poor air circulation or avian respiratory disease. Although a major problem in the 1980s, and still a problem today when circumstances combine, by the mid-1990s PHS and SDS were brought under control by a combination of family-based genetic selection against the condition, and careful management minimizing the external factors that contribute to hypoxia.

### **Stage 3: Combined Selection for JWfA, FCR, and Proportion of Breast Meat (PBM)**

From the early 1970s, selection of birds with a less-pronounced keel bone was a part of broiler selection,

### **Box 3 Etiology of Pulmonary Hypertension Syndrome and Sudden Death Syndrome**

The causal sequence leading from selection for JWfA and FCR to PHS is thought to be as follows: Selection for rapid juvenile growth rate increases metabolic rate for growth resulting in increased need for oxygen in tissues. At the same time, selection for decreased FCR caused the relevant regulatory, circulatory, and hormonal systems to reduce oxygen consumption by the tissues resulting in effective hypothyroidism, and also reduced the proportion of oxygen-supplying organs (lung and heart) relative to oxygen-demanding tissue (muscle). The combination of these two factors results in tissue hypoxia that leads, via the kidney and erythropoietin, to increased production of red blood cells that increases blood viscosity, and also to constriction of pulmonary arterioles to ensure blood flow to all parts of the lungs so as to increase pulmonary oxygen exchange. The result of these two factors was increased lung arteriole pressure leading to increased workload on the heart and hypertrophy of the right (pulmonary) ventricle. This in turn resulted in incomplete closure of the right arterial valve with consequent backup of blood pressure to the hepatic and portal veins generating pulmonary hypertension and impaired uptake of fluid by the lymphatic system. Pulmonary hypertension results in fluid leakage and accumulation of fluid in the lungs (hypertensive lung syndrome) and in the abdominal and pericardial cavities (Ascites), and inability of the heart to supply body oxygen needs through left ventricle heart failure. Any of these final syndromes can result in death of the chick from PHS.

The detailed etiology of SDS differs [34, 35]. Electrocardiograms of broilers in the last stages of SDS show the immediate cause of death in SDS to be acute cardiac arrhythmia terminating in ventricular fibrillation. It is thought that under conditions of hypoxia the myocardium may become hyperirritable, serving as a secondary pacemaker and interfering with normal cardiac rhythms. Indeed, studies of rapidly growing broilers show that they manifest a high rate of cardiac arrhythmia under “normal” high growth-rate conditions, and are highly susceptible to stress-induced cardiac arrhythmia. Thus, under normal high growth-rate conditions, any additional stress may be sufficient to tip the heart from simple arrhythmia to ventricular fibrillation and sudden death [32, 33].

with primary aim of producing a finished product having a fuller appearance that was more attractive to the consumer. This had a secondary result of selecting birds with more breast meat. With an increasing percentage of birds sold as individual parts, a major price differential between chicken breast “white” meat and chicken leg and thigh “dark” meat became apparent. This meant that the income of the broiler producer was primarily a function of the amount of breast meat produced. In the mid-1980s, realization of this economic fact by the breeders led to the addition of specific selection for a high proportion of breast meat in the broiler carcass as a third major breeding goal in addition to JWfA and FCR. This continues to the present time.

#### **Secondary Effect of Selection for JWfA and PBW (XI): Heat Stress (Again)**

The major secondary effect of selection for PBW in the broiler was an additional increase in heat stress effects, due to the increased metabolic load induced by the high metabolic cost of muscle mass synthesis. As a result, the modern broiler could no longer cope with even small deviations from the neutral temperature zone without a significant loss of meat yield. It became clear that the typical broiler-rearing shed of minimal structure with open sides plus fan-ventilation could no longer deal successfully with the enormous metabolic heat produced by the modern broiler flocks. This led the industry to adopt a closed shed and a shift to “tunnel ventilation” with further option of cooling the air by drip technology. During the 2000s, “tunnel ventilation” with “drip cooling pads” for broilers became an obligatory feature of broiler rearing.

#### **Secondary Effect of Selection for PBM (XII): Quasi-EODES at Entry to Lay**

A secondary effect of the selection for PBM was the appearance of a new “quasi-EODES” syndrome at the onset of lay in the broiler breeder female parent, with symptoms very similar to classical EODES but to a less-pronounced degree [36]. This “quasi-EODES” syndrome was characterized by high mortality at entering lay, caused mainly by cloacal prolapse, internal lay, and inflammation of the oviduct. For the surviving birds, the quasi-EODES syndrome has marked negative effects

on peak lay and eggshell quality resulting in a lower proportion of hatching eggs out of all eggs and lower hatch of healthy chicks from fertile eggs. The overall result is a marked reduction in chick production. Practical experience of hatchery flock managers, supported by later research revealed that the quasi-EODES syndrome is induced by even minor “overfeeding” during the critical weeks of forced sexual maturation. The modern female breeder became very sensitive to even a slight deviation from optimal feed restriction.

Two factors appear to have combined to generate the quasi-EODES syndrome. The first, is the continued increase in the degree of feed restriction during the pullet growth period, needed for preventing EODES, going, as noted above, from about 70% of *ad libitum* consumption during the 1970s to about 35% of *ad libitum* consumption at present. The second is the continued increase in threshold weight for onset of sexual maturity of the female broiler breeder, as a direct result of the selection for higher JWfA and PBM. For example, recommended 24-week body weight for out-of-season “Cobb 500” female broiler breeders increased from 2700 g in 1987 to 3160 g in 2005 [37] with a higher proportion of body mass as muscle tissue. Thus, the actual gap between the pullet lean mass attained at the point where light and feed stimulation is initiated, and the threshold lean mass required for entering lay has progressively increased, and with it the time required from initiation of light and feed stimulation to actual onset of lay. Attempting to cover this gap more rapidly by increasing feeding levels exceeds the ability of the pullet to rebuild breast mass and results in a surplus of nutrients causing the ovary to behave as if fed *ad libitum* and leading to development of quasi-EODES. Thus, the modern broiler breeder pullet entering lay is delicately balanced at a feeding schedule that will induce lay as rapidly as possible, yet will not induce quasi-EODES. Consequently, the bird is very unforgiving of any deviation from optimum feeding schedule in the direction of overfeeding.

#### **Secondary Effect of Selection for PBW (XIII): Hatchability Problems (Again)**

A further secondary effect of the selection for increased proportion of breast meat was a reemergence of hatchability problems. These appeared even though

continued selection for rapid JWfA did not result in appreciable further increase in egg weight. Stability of egg weight was achieved due to increasing feed restriction of the broiler mother targeted specifically at controlling egg weight (which responds very rapidly to feed restriction) and to conscious selection against excessive egg size by the breeders. Thus, the situation was unchanged throughout the 1990s. However, the intense selection for increased muscle mass that began in the mid-1980s resulted in an increase in the proportion of muscle mass in the late developing embryo. As noted, the metabolic energy expended in development of muscle is greater than for other body tissues, because of the high metabolic cost of protein synthesis. Hence, this resulted in a further increase in metabolic heat production by the late embryo [38–41]. This additional excess heat was beyond the capacity of multistage incubators, no matter how carefully managed. Consequently, hatchability began to fall again, owing to increased late mortality and decreased yolk uptake leading to an increased proportion of weak chicks, and more grade B chicks [42], and negative effects on broiler bone development and leg health [43, 44].

To meet this challenge, the industry began shifting to “single-stage” incubators [40]. In these incubators, all eggs from flocks of about the same age and same average egg weight are loaded into the incubator on the same day. They are then incubated together until they hatch. Thus, they are all at the same stage of development at any given time (hence the name “Single-Stage” incubator). This makes it possible to better adjust incubation temperature, according to the stage of the embryos [42] – somewhat warmer at the early embryo stages, when the embryo requires external heat to reach desired internal temperature, and cooler at the later embryo stages, when the embryo needs to rid itself of excess heat. Similarly, humidity can also be adjusted so as to achieve optimal egg water loss across the incubation period [41]. This has proven to be a satisfactory solution, and hatchability and chick quality have improved accordingly.

#### Secondary Effect of Selection for PBW (XIV): Reduced Locomotor Activity

The very large breast of the modern broiler chick results in an animal that is unbalanced between breast weight

and leg and thigh muscles [45]. Consequently, it is difficult for the animal to move, and locomotor activity is much reduced. Toward the end of the growth period, the broiler chick spends over 90% of its time sitting and lying, resting its overdeveloped breast on the litter, and basically moving only to eat and drink. Selection for reduced FCR may also have contributed to the reduction in locomotor activity, since this is one way to reduce metabolic expenditure of the animal. As a result, the animal is very susceptible to contact dermatitis, including footpad lesions and breast blisters. These conditions can be well controlled by meticulous attention to litter quality, specifically litter temperature and humidity, but are exacerbated and can reach high levels when litter quality is poor.

### The Role of Breeding and Management in the Improvement in Performance of the Modern Broiler Chicken

#### Achievements of the Modern Commercial Broiler with Respect to the Three Main Production Traits and the Associated Secondary Traits

The extraordinary improvement in broiler performance for the three main production traits (JWfA, FCR, and PBW) over the past 60 years is best captured by the following table (Table 1), which compares the performance (average of males and females) of representative commercial broiler stocks of 1957, 1991, and 2001.

**Poultry Breeding. Table 1** Comparison of performance of broiler stocks of 1957, 1991, and 2001<sup>a</sup>

Trait	1957	1991	2001
Age at sale (d)	84	42	42
Live weight at sale (g)	1646	2132	2672
Feed conversion ratio	3.26	2.04	1.63
Breast meat (%)	12.9	15.0	20.0
Carcass fat (%)	14.7	14.1	13.7
Heart+lungs (%)	1.242	ND	1.023
Mortality to market (%)	2.52	9.70	3.57

<sup>a</sup>Based on [6–9]; data for 1957 average of [6 and 8 and 7 and 9] for birds fed the commercial feed formulations of 1991 [6, 7] and 2001 [8, 9].

It is evident that in the period 1957–2001 enormous improvement has been achieved in the three major production traits. Days to market weight and FCR have been reduced by half, while JWfA has almost doubled, and PBW has increased by almost 30%. Improvement in all traits was continuous, with no indication of reduced rate of gains in the latter decade.

When compared using the feed formulations of 1957, gains in JWfA and FCR were only a bit less. There is no doubt therefore, that genetic selection by the primary breeders was by far the main driving force leading to the gains in these traits. The increase in PBM from 1991 to 2001 is particularly impressive and can be attributed almost completely to the attention devoted to this trait by the primary breeder beginning in late 1980s. However, JWfA and FCR also increased strongly during this period showing the ability of the breeder to deal simultaneously and effectively with the three major production traits.

Along with the phenomenal improvement in production traits, the broiler industry was able to cope successfully with all of the deleterious secondary effects detailed in the preceding sections that accompanied the genetic gains in the three main production traits. Through a combination of genetic selection on the one hand, and continual adjustment and innovation in broiler and breeder flock management and physical facilities on the other, the industry succeeded in maintaining male and female reproductive performance at very high levels, while keeping losses due to heat sensitivity, skeletal problems, and metabolic disorders at very low levels.

### **Relative Contribution of Genetics and Management to Modern Broiler Performance**

The relative contribution of genetics and management to these achievements depends on the specific trait. In particular, it seems clear that the shift to progressively more active ventilation culminating in the tunnel ventilation with drip cooling in general use at present was the main means for coping with the heat sensitivity of the modern broiler. Similarly, changes in incubator management and the more recent shift from multistage to single-stage incubators together with flock management to reduce egg weight were the main factors maintaining high hatchability rates. However, selection

played some role in both cases, since families that display low hatchability and high sensitivity to heat stress are generally culled from the breeding population.

With respect to reproductive performance, feed restriction during pullet growth and the hen laying period, and dark-out housing of the parent stock pullets with stimulatory lighting and feeding at entry into lay, were critical management factors contributing to maintaining reproductive performance in male and female broiler parent stock. However, selection for reproductive performance within this overall management scheme played an important secondary role, as attested by the appreciable differences between genetic stocks in reproductive performance under optimal feeding and lighting schedules. In the same way, while appropriate management is essential to control skeletal problems and metabolic disorders, the effective selection applied to these conditions appears to have been crucial in reducing their incidence in well-managed flocks to present-day low single-digit proportions.

Thus, although the improvement in production traits of the modern broiler compared to its original founder populations can be attributed almost entirely to genetic improvement achieved by selection, the bird that results is able to achieve optimal production and functional performance as broiler and parent stock primarily due to major modifications in broiler housing, and in parent flock and incubator management coupled with a greater or lesser contribution by selection, depending on the trait. Be that as it may, one can only stand in awe at the combined ability and synergistic work of the major players in the industry to overcome the challenges as they arose while maintaining continual progress in overall efficiency of the broiler enterprise.

### **Breeding Methods for Genetic Improvement**

In large part, this is due to the fact that the methods practiced by the primary broiler breeders increased in sophistication in accord with the complexity of the breeding challenges that they faced [46]. In the first stage of broiler genetic improvement, JWfA was the main trait under selection, with a minor goal of improved breast conformation. This was achieved by simple two-step tandem mass selection, the first step

being selection for JWfA followed by a second step of selection for breast conformation. Flocks were not pedigreed at this stage. With the addition of reproductive performance, FCR and reduction of adiposity and skeletal problems as breeding objectives in the second stage of broiler genetic improvement, simple mass selection was no longer effective and multiple-stage tandem selection no longer feasible. To meet the new challenges, breeders changed to fully pedigreed flocks to provide information on reproductive performance, FCR, and skeletal problems, and to index selection based on individual and close relatives to combine information on the various traits in an optimal manner. Finally, in the third stage of broiler genetic improvement, with the addition of PBM and reduction of metabolic disorders to the breeding objectives, breeders adopted BLUP and Individual Animal Model statistical methodologies for estimating breeding values for the various traits under selection, based on individual data for the entire current and past population. At the same time, the primary breeders adopted phenotyping procedures that increased the completeness and accuracy with which the traits of importance were evaluated. For example, individual skeletal problems are now assessed by x-ray scanning of bird joints and individual cardiovascular function is assessed by blood oximeter machine. At present, over 50 traits are recorded on the individuals in the breeding nuclei, about half of these relating to various aspects of broiler health.

At all stages, breeding nuclei were kept very large in order to avoid inbreeding and maintain genetic variation. With the advent of genome-wide procedures for estimation of breeding values [1–3], all of the major primary breeders have instituted in-house experimental studies to evaluate the use of these methodologies in their practical breeding programs. Although very promising in initial studies, the results in practice in commercial breeding nuclei remain to be seen.

### **Accounting for the Long-Term Continuous Response to Selection and for the Punctuated and Coordinated Appearance of Secondary Effects**

Commercial breeding nuclei have been under intense and effective directional selection for JWfA for 60

generations, for FCR for 40 generations, and for PBM for 25 generations. Two remarkable features characterize the response to this selection. The first and most striking is that the direct response of the target traits to selection has been positive and continuous for all three traits, with no indications of having reached a selection plateau for any of them. This continued response to selection requires explanation. Since selection acts on all relevant loci at the same time, the intense directional selection for the target traits should have exhausted the existing genetic variation in the original dual-purpose chicken founder populations rather rapidly, leading to selection plateaus for the various traits. The second remarkable feature is the rather sudden manifestation of the secondary effects of the response to selection at about the same time period in the stock of all leading breeders, even though their respective breeding nuclei are effectively isolated from one another. This “punctuated” (i.e., “relatively sudden”) and “coordinated” (i.e., relatively universal and synchronous) manifestation of the secondary effects requires explanation. If due to genetic correlations or pleiotropic effects of the initial genetic variation of the founder populations, they would be expected to manifest in a more linear and gradual manner. If due to pleiotropic effects of new mutations, they would be expected to manifest in a more sporadic manner, affecting only stock of one or two of the various breeders.

### **Sources of Genetic Variation for the Long Continued Response to Selection**

Based on classical population genetics, three sources can be envisaged with respect to sources of genetic variation to serve as a substrate for selection. The first is the genetic variation present in the original founder populations of the broiler stocks. Recent research in mapping of the loci responsible for genetic variation in production traits in farm animals (the so-called Quantitative Trait Loci or QTL) indicates that tens of QTL may contribute to genetic variation of such traits. This genetic variation present in the founder populations would have provided the basis for the first decades of response. The second source comprises rare alleles with positive effects on the target traits that were present at low frequency in the founder populations. The primary breeders maintain large breeding nuclei, and these

could have included an appreciable number of such loci. In the early generations, because of their low frequency these loci would not have contributed materially to genetic variation in the target traits, or to the response to selection. However, after several decades of intense selection, their frequency would have increased to the point where they could serve as useful targets for the middle stages of selection. Finally, new mutations or rare recombinants would supply a steady stream of new genetic variation that would come into play in the later stages of selection.

### **Punctuated and Coordinated Appearance of Deleterious Secondary Effects**

Similarly, a number of explanations can be offered within classical frameworks, for the punctuated and coordinated appearance of deleterious secondary effects. Some of these effects may be due to non-linear secondary effects of selection interacting with downstream threshold effects. For example, adiposity can be presumed to have increased more or less linearly with selection for JWfA, but the effect on egg production was virtually nil for the first 10–15 generations of selection, and then suddenly became very severe – in the few years just before the introduction of feed restriction in the late 1960s, yearly egg production had fallen by almost 100 eggs per laying hen. Such effects would be expected to manifest in flocks of all breeders at about the same time, as their breeding populations reached more or less the same degree of expression of the primary traits. Other secondary effects may be due to the introduction of new primary target traits. Thus, an entire group of new secondary effects manifested following the introduction of FCR, and again following the introduction of PBM as targets for selection. Since the primary breeders introduced the same new target traits for selection at more or less the same time, these secondary effects would be expected to manifest in flock of all breeders at about the same time.

### **Difficulties with the Explanations Within the Classical Framework of Population Genetics**

Rare mutations are often those that are held at low frequencies because of deleterious pleiotropic effects on fitness. Similarly, new mutations generally affect

numerous traits, with deleterious fitness effects on one or more of them. Intense selection can increase the frequency of genes with strong positive effect on the target trait in spite of negative effects on fitness, but in a long-term selection program this should result in reduced overall functional fitness of the population. Furthermore, because of their low frequency in the founder populations, rare alleles would not distribute uniformly among different breeding populations derived from the same founders. The same would certainly hold for the genes that were affected by mutation. Consequently, fitness traits should be differentially affected in different breeding populations, according to the specific low-frequency genes that were inherited or mutated in each population. This does not appear to have been the case. Similarly, although in almost all cases it was possible to trace in retrospect a causal chain leading from selection for the primary trait to the manifestation of the secondary trait it is somewhat unexpected that the specific form of these manifestations (e.g., PHS, SDS, quasi-EODES, reduced photosensitivity) was the same in stocks of all breeders. If much genetic variation is due to initially rare or mutated loci that are necessarily specific and different in different breeding populations, specific manifestations of secondary traits would be expected to differ as well.

### **Selection-Induced Genetic Variation (SIGV)**

The difficulties with the explanations proposed within the classical framework for long-term response to selection and for punctuated and coordinated appearance of the secondary manifestations have led to a recent proposal for a new previously unrecognized source of genetic variation in ongoing selection programs: selection-induced genetic variation (SIGV) [47]. The SIGV hypothesis is based on the observation that the individual QTL determining genetic variation in production traits often exhibit strong epistatic interactions with one another and with the genetic background. Consequently, a segregating locus with zero effect on the target trait in one genetic background might have a strong effect on the trait in a different genetic background.

The intense continued directional selection, changes the frequency of many alleles in many loci and by means of “hitchhiking effects” of their closely



linked neighboring genes as well. This changes the genetic architecture of a population to a marked degree. Thus, some loci that were neutral in their effect on the target trait at the beginning of the selection program, may transform into sources of genetic variation as the selection program progressively modifies the genetic architecture of the population. As the positive alleles at these loci are caught up and brought to high frequency by the ongoing selection, the genetic architecture of the population becomes modified again, and new previously neutral loci come into play as sources of genetic variation. In this way, the population under selection would continually generate new genetic variation to serve as a substrate for the next phase of the directional selection process, with no obvious limit in place.

Since the loci involved in the SIGV process are loci that were present at appreciable frequencies in the original founder populations, they would be present in all breeding stocks, and would be expected to come into play at about the same stage in the selection process in the individual flocks. Thus, genetic progress in the various breeder populations would not depend on rare alleles or new mutations, which would be specific to each population, but on segregating loci already present in the population at useful frequencies, that come into play in a programmed manner, as the genetic architecture of the population changes. On this hypothesis, the genetic architecture of all breeding populations under the same general selection regime would be similar at all stages of the selection process, and hence the specific form of any secondary effects would also be similar in the different populations. Strong secondary effects on fitness would not be expected, since the loci involved are loci historically present in the populations, and hence have survived the screening effect of natural selection.

Thus, the SIGV hypothesis provides plausible explanations both for the long-term response to selection and for the punctuated and coordinated appearance of secondary effects. Indeed, recent experimental studies analyzing crosses between layers and broilers, and between two-way single-trait selection lines for JWfA, provide significant experimental support for this hypothesis [48, 49]. The SIGV hypothesis emphasizes the importance of epistatic interactions in genetic variation and encourages the development of statistical and genomic methodologies that can exploit these interactions.

## Genetic Improvement of Broiler Production Traits and Broiler Welfare

In the previous sections, it was seen that the modern broiler is greatly modified developmentally and physiologically from the original dual-purpose breeds from which it was developed. Furthermore, to achieve optimal economic performance the modern broiler and its parent flocks must be reared, managed, and housed under conditions very different from the husbandry conditions of the founder parent lines. Various aspects of these genetic and management changes impinge on animal welfare, in the sense that they can lead to distress or pain of the individual. These can be considered in two main categories. Those where the distressful situation is unavoidable, and those where the distressful situation is avoidable under good management, but manifests under less than optimal management. There is a third category, which considers aspects where the bird is comfortable and not in distress, but where it is thought that its life might be more joyful if living under different conditions, for example, on free range or scrounging for food in the jungle or farmyard, instead of in a closed shed. This latter raises more profound issues of the morality and ethics of animal agriculture in general and industrial animal agriculture in particular, and will not be considered here.

### Unavoidable Distress

The negative effect of excess adiposity on reproductive performance and survival of the modern broiler parent female or male is dramatic and extreme. Consequently, stringent feed restriction of the parent flock pullets and males during the rearing and laying period is essential, if the birds are to achieve optimal reproductive performance. This appears to be the only aspect of the broiler enterprise that causes distress, but is unavoidable. An animal that is fed only 35% of its *ad libitum* consumption, and can be seen to eat ravenously when feed is available, must surely be feeling hunger pangs. The period of stringent feed restriction is rather short, however, from about 4–14 weeks of age, and while it can be presumed to be uncomfortable, does not affect the overall health of the individuals in any way. Indeed, allowing the animal to eat *ad libitum* and become grossly obese would probably entail much more distress over the entire life span of the animal including

mortality. The much milder feed restriction that is in effect during lay would not seem to pose a welfare problem, as it is ordinarily sufficiently generous to even allow for some increase in body weight over the course of the laying period.

### **Avoidable Distress**

A review of the preceding sections, however, reveals a considerable list of conditions to which the modern broiler is susceptible due to its changed genetic constitution. These include: quasi-EODES at entry into lay that can result in death from cloacal prolapse or from internal lay; heat stress, skeletal abnormalities, PHS which often manifests as ascites, SDS, and various forms of contact dermatitis. SDS, in which death of the animal occurs within less than a minute from the onset of first symptoms of distress, would seem to occur too rapidly to be a source of significant pain or distress. But the other conditions can be presumed to be moderately to severely painful, according to their respective natures and degree of severity. As emphasized in the previous sections, thanks to the selective work of the breeder on the one hand, and innovative management of the industry on the other, when broiler commercial or parent flocks are managed with care and attention to detail, all of the above can be kept at very low single-digit levels [46, 50].

### **Importance of Stocking Density**

In this context, stocking density, an aspect of management that was not previously emphasized becomes important [45]. Stocking density interacts with the susceptibilities of the modern broiler in two ways: increasing heat stress and decreasing litter quality with consequent increase in the incidence of contact dermatitis. When stocking density is high, the bodies of the chicks cover the entire floor area, forming a barrier between the litter surface and the ventilated area of the shed. In addition, higher stocking density provides additional nitrogen and moisture to the litter increasing heat production due to microbial growth and metabolism. Taken together, these two factors can generate an appreciable differential between temperature at broiler level and temperature a meter above the floor. High stocking density also limits air movement and increases radiant heat transfer between birds, further increasing heat

stress. The increase in litter temperature, humidity, and ammonia concentration resulting from high stocking density also interacts with the limited locomotor activity of the broilers to increase the incidence of contact dermatitis. Accordingly, proper control of stocking density is essential for optimal broiler welfare, and may be less than the economic optimum.

### **From Farm to Market**

The handling of the bird on its final path from shed to slaughterhouse is another case in point. Because of the young age of the broiler at market weight, and its minimal locomotor activity, the long bones of the legs and wings are fragile relative to body weight. Moving the birds from the grow-out shed to the slaughter line is low-wage work with high personnel turnover and great pressure to get the work done with minimal labor costs. Consequently, the birds may be roughly handled at this time resulting in leg and wing bone breakage when they are captured and moved from the pen to the transport crates, and again from the transport crates to the slaughter line [51, 52]. Such breakage results in direct losses to the farmer and hence normally efforts are expended to reduce this to a minimum through mechanization, proper design of facilities, and training of personnel [53].

### **Welfare and Society**

Apparent from the above paragraphs is the critical dependence of broiler welfare on meticulous flock and personnel management. Due to the work of the breeder, with proper management the susceptibilities of the bird remain latent. But there is little room for error. The bird is unforgiving of even slight deviations from optimal management, and will react to such deviations by manifesting one or other of its innate susceptibilities. The modern chicks that manifest any of the conditions above are a net loss to the farmer, and management conditions conducive to such manifestation are often inimical to optimal growth and reproduction. Consequently, for the most part, the welfare of the chick and the economic interests of the farmer correspond. However, achieving optimal management has its costs as well, and economic maximization may result in a management program that is somewhat less than optimal for the bird.

Society as a whole has benefited greatly from the improved production characteristics of the modern broiler, which has added a low-cost highly nutritious staple to the menu at minimal environmental footprint. Along with these benefits comes responsibility for the welfare of the bird in those instances where economic interests of the farmer and the quality of life of the broiler are not perfectly aligned. This is true throughout the broiler enterprise; nowhere more than in the final hours of the chick's life. Happily, society is accepting these responsibilities by designing facilities based on understanding of livestock behavior [53], and by defining acceptable production standards, and embodying these in appropriate regulatory legislation [54, 55].

### Future Directions

One stands in awe of the magnificent achievements of the primary broiler breeders who have succeeded in improving the production characteristics of the modern broiler manifold relative to the initial dual-purpose founder populations, while at the same time dealing successfully with the many secondary functional problems that arose along the way, maintaining high reproductive performance together with very low incidence of health problems. This was not the achievement of the breeder working alone, of course, and at all stages involved combined contributions by all components of the industry in devising novel management programs and physical facilities such as feed restriction and stimulatory photoperiods, dark-out housing, and tunnel ventilation that enabled optimal functional performance of the birds.

### The Resource Allocation Model

These results are all the more remarkable, because they differ from what would be expected on the widely accepted resource allocation model for interpreting the effects of selection for production traits on functional traits [56, 57]. The basic assumption of this model is that under the conditions of intensive agriculture, the energy resources available to an animal are fixed. The animal must then allocate these resources to maintenance, to production traits including growth, and to functional traits including reproduction, regulation of physiology and metabolism, and defense against pathogens and toxins. Selection for production

traits diverts resources preferentially to these traits, leaving less for the functional traits. Consequently, functional traits are expected to suffer. Indeed, many experimental studies show that production traits suffer when resources are preferentially allocated genetically or experimentally to functional traits, and functional traits suffer when resources are preferentially allocated to production traits. Amazingly, the broiler industry appears to have been able to circumvent this negative correlation, achieving high levels in both production and functional traits.

### Management and Biological Costs of the Performance Achievements of the Broiler Industry

More detailed consideration, however, shows these achievements have incurred two costs. The first is the cost of more sophisticated physical facilities and complexity of management. These include sheds with tunnel ventilation and drip cooling for broiler rearing; dark-out houses for pullet rearing; feed restriction, single-stage incubators; and all of the many management and nutritional modifications required for optimum performance. The second cost is the apparent loss of buffering capacity of the animal regulatory systems, resulting in ever-increasing fragility or sensitivity of the animal to environmental challenges. This sensitivity expresses itself as reduced production and functional performance, and increased incidence of health problems as a consequence of even minor deviations from optimum management and environment. The result is not only loss of economic value, but also harmful impact on animal welfare. Because of the high metabolic cost of protein synthesis, a provocative hypothesis suggests that selection for productivity in general, and FCR in particular may reduce protein turnover rate. Since protein synthesis is essential to meet new conditions, this may be the underlying cause for the general loss in buffering capacity [58]. The primary broiler breeders today are attempting to decrease the sensitivity of their populations to environmental insults by testing under multiple environments and selecting for "robustness", as they would select for any other positive health trait. Will they succeed? Or will the future broiler require an increasingly narrow and complex multidimensional management path for optimal performance, with steep losses outside of this path?

The coming years will tell, with strong implications for the broiler industry, and for animal breeding in general.

### Conclusions and Interpretation

Be that as it may, the broiler experience teaches that deleterious effects on functional traits are almost certain to arise in the course of selection, and can take unexpected forms. The industry, whatever the species, must be alert to this possibility, so as to seek solutions as early as possible. The broiler experience also teaches that the breeder and the industry acting in concert can indeed find these solutions to these problems, enabling a joint genetic-management program that provides very high productive performance together with high functional performance within a framework of innovative physical facilities and increased management complexity. Based on past experience, therefore, as animal breeding enters a new genomic phase with greatly accelerated improvement in production traits, the future can be looked at with appropriate confidence in the ability of the breeder and industry working together to achieve sustained genetic improvement in production traits while maintaining high levels of functional and health performance.

### Abbreviations

BLUP	Best linear unbiased predictor
DNA	Deoxyribonucleic acid
EODES	Erratic oviposition defective egg syndrome
FCR	Feed conversion ratio
JWfA	Juvenile weight for age
PBM	Proportion of breast meat
PHS	Pulmonary hypertension syndrome
QTL	Quantitative trait locus
RNA	Ribonucleic acid
SDS	Sudden death syndrome
SIGV	Selection-induced genetic variation

### Bibliography

#### Primary Literature

- Hayes BJ, Bowman PJ, Chamberlain J, Goddard ME (2009) Invited review: genomic selection in dairy cattle: progress and challenges. *J Dairy Sci* 92:433–443
- Tier B (2010) Editorial: genomic selection: promise and propriety. *J Anim Breed Genet* 127:169–170
- Meuwissen T, Goddard M (2010) Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics* 185:623–631
- Hartmann W (1989) From Mendel to multi-national in poultry breeding. *Worlds Poult Sci J* 45:5–26
- Schmidt CJ, Persia ME, Feierstein E, Kingham B, Saylor WW (2009) Comparison of a modern broiler line and heritage line unselected since the 1950's. *Poult Sci* 88:2610–2619
- Havenstein GB, Ferket PR, Scheideler SE, Rives DV (1994) Carcass composition and yield of 1957 vs 1991 broiler when fed "typical" 1957 and 1991 broiler diets. *Poult Sci* 73:1795–1804
- Havenstein GB, Ferket PR, Scheideler SE, Larson BT (1994) Growth, livability and feed conversion of 1957 vs 1991 broiler when fed "typical" 1957 and 1991 broiler diets. *Poult Sci* 73:1785–1794
- Havenstein GB, Ferket PR, Qureshi MA (2003) Growth, liveability and feed conversion of 1957 vs 2001 broilers when fed representative 1957 and 2001 broiler diets. *Poult Sci* 82:1500–1508
- Havenstein GB, Ferket PR, Qureshi MA (2003) Carcass composition and yield of 1957 versus 2001 broilers when fed representative 1957 and 2001 broiler diets. *Poult Sci* 82:1509–1518
- F.A.O. Food outlook global market analysis (2010). <http://www.fao.org/docrep/012/ak349e/ak349e00.pdf>. Accessed April 2010
- Anonymous (1996) The history of Hubbard Farms. Hubbard Farms, Walpole
- National Chicken Council. US chicken industry history. <http://www.nationalchickencouncil.com/>. Accessed April 2010
- Emmerson DA (1997) Commercial approaches to genetic selection for growth and feed conversion in domestic poultry. *Poult Sci* 76:1121–1125
- Siegel PB, Wolford JH (2003) A review of some results of selection for juvenile body weight in chickens. *Poult Sci J* 40:81–91
- Soller M, Eitan Y (1984) Why does selection for live weight gain increase fat deposition? A model. *Worlds Poult Sci J* 40:5–9
- Wikipedia The metabolic syndrome. [http://en.wikipedia.org/wiki/Metabolic\\_syndrome](http://en.wikipedia.org/wiki/Metabolic_syndrome). Accessed April 2010
- Jaap RG, Muir FV (1968) Erratic oviposition and egg defects in broiler type pullets. *Poult Sci* 47:417–423
- Katanbaf MN, Dunnington EA, Siegel PB (1989) Restricted feeding in early and late feathering chickens 2. Reproductive response. *Poult Sci* 68:352–358
- Robinson FE, Robinson NA, Scott TA (1991) Reproductive performance, growth rate and body composition of full fed versus feed restricted broiler breeder hens. *Can J Anim Sci* 71:549–556
- Hocking PM, Waddington D, Walker MA, Gilbert AB (1989) Control of the development of the ovarian follicle hierarchy in broiler breeder pullets by food restriction during rearing. *Br Poult Sci* 30:161–174
- Nir I, Nitsan Z, Dror Y, Shapira N (1978) Influence of over feeding on growth, obesity and intestinal tract in young chicken of light and heavy breeds. *Br J Nutr* 39:27–35

22. Boerjan M (2004) Genetic progress inspires changes in incubator technology. *World Poul* 5:16–17
23. Leach RM (1992) Leg weakness in broilers – complex situation involving many factors. *Poult Dig* 3:24–27
24. Oviedo-Rondon EO, Ferket PR, Havenstein GB (2006) Understanding long bone development in broilers and turkeys. *Avian Poult Biol Rev* 17(3):77–88
25. Nicholson D (1998) Research: is it the broiler industry's partner into the new millennium. *World's Poult Sci J* 54:271–278
26. Tomas FM, Pym RA, Johnson RJ (1991) Muscle protein turnover in chickens selected for increased growth rate, food consumption or efficiency of food utilization: effects of genotype and relationship to plasma IGF-I and growth hormone. *Br Poult Sci* 32:363–376
27. Skinner-Noble DO, Teeter RG (2003) Components of feed efficiency in broiler breeding stocks: energetics, performance, carcass composition, metabolism and body temperature. *Poult Sci* 82:1080–1090
28. Lewis P (2009) Proper lighting for broiler and turkey breeders. *World Poul* 25(12):18–21
29. Eitan Y, Soller M (1994) Selection for high and low threshold body weight at first egg in broiler strain females. 4. Photoperiodic drive in the selection lines and in commercial layers and broiler breeders. *Poult Sci* 73:769–780
30. Teeter RG (1994) Optimizing production of heat stressed broilers. *Poult Dig* 53:10–27
31. Yahav S (2010) Alleviating heat stress in domestic fowl: different strategies. *Worlds Poult Sci J* 65:719–732
32. Leeson S (2007) Metabolic challenges: past, present and future. *J Appl Poult Res* 16:121–125
33. Decuyper E, Buyse J, Buys N (2000) Ascites in broiler chickens: exogenous and endogenous structural and functional causal factors. *Worlds Poult Sci J* 56:367–377
34. Moghadam HK, McMillan I, Chambers JR, Julian RJ, Tranchant CC (2005) Heritability of sudden death syndrome and its associated correlations to ascites and body weight in broilers. *Br Poult Sci* 46:54–57
35. Olkowski AA, Classen HL (1997) Malignant ventricular dysrhythmia in broiler chickens dying of sudden death syndrome. *Vet Rec* 140:177–179
36. Eitan Y, Soller M (2009) Problems associated with broiler breeder entry into lay: a review and hypothesis. *Worlds Poult Sci J* 65:641–648
37. Cobb-Vantress (2005, 2008) Cobb 500 breeder management guide. Cobb-Vantress, Arkansas
38. Taylor G (1999) High yield breeds require special incubation. *World Poul* 15(3):27–29
39. Hill D (2000) Multistage incubation and yield genetics. *Poult Dig* 10–11:14–20
40. Boerjan M (2004) Single stage incubation is the most natural choice. *World Poul* 20(7):18–20
41. Tona K, Bamelis F, Coucke W, Bruggeman V, Decuyper E (2001) Relation between Broiler Breeder's age and egg weight loss and embryonic mortality during incubation in large-scale conditions. *J App Poult Res* 10:221–227
42. Meijerhof R (2002) Incubation by embryo temperature. *World Poul* 18(6):36–37
43. Oviedo-Rondon EO, Wineland MJ, Funderburk S, Small J, Cutchin H, Mann M (2009) Incubation conditions affect leg health in large high yield broilers. *J Appl Poult Res* 18:640–646
44. Oviedo-Rondon EO, Wineland MJ, Small J, Cutchin H, McElroy A, Barry A, Martin S (2009) Effect of incubation temperatures and chick transportation conditions on bone development and leg health. *J Appl Poult Res* 18: 671–678
45. Bessei W (2006) Welfare of broilers: a review. *Worlds Poult Sci J* 62:455–466
46. Katanbaf MN, Hardiman JW (2010) Primary broiler breeding – striking a balance between economic and well-being traits. *Poult Sci* 89:822–824
47. Eitan Y, Soller M (2004) Selection induced genetic variation: a new model to explain direct and indirect effects of sixty years of commercial selection for juvenile growth rate in broiler chickens. In: Wasser SP (ed) *Evolutionary theory and processes: modern horizons. Papers in honor of Eviatar Nevo*. Kluwer, Dordrecht, pp 153–176
48. Deeb N, Lamont SJ (2002) Genetic architecture of growth and body composition in unique chicken populations. *J Hered* 93:107–118
49. Carlborg O, Jacobsson L, Ahgren P, Siegel P, Andersson L (2006) Epistasis and the release of genetic variation during long-term selection. *Nat Genet* 38:418–420
50. Flock DK, Laughlin KE, Bentley J (2005) Minimizing losses in poultry breeding and production: how breeding companies contribute to poultry welfare. *Worlds Poult Sci J* 61:227–237
51. Nicol CJ, Scott GB (1990) Pre-slaughter handling and transport of broiler chickens. *Appl Anim Behav Sci* 28:57–73
52. Grandin T, Johnson C (2009) *Animals make us human - Creating the best life for animals*. Houghton Mifflin Harcourt, Boston/New York
53. Grandin T, Deesing M (2008) *Humane livestock handling: understanding livestock behavior and building facilities for healthier animals*. Storey Publ., North Adams
54. Shane S (2009) Broiler welfare symposium – The balance between producers' and consumers' standards. *World Poul* 25(12):32–34
55. Blokhuis HJ (2004) Recent developments in European and international welfare regulations. *Worlds Poult Sci J* 60:469–477
56. Gross WB, Siegel PB (1997) Why some get sick. *J Appl Poult Res* 6:453–460
57. Siegel PB, Honaker CF (2009) Impact of genetic selection for growth and immunity on resource allocations. *J Appl Poult Res* 18:125–130
58. Rauw WM, Knap PW, Varona L, Gomez-Raya L, Noguera J (2002) Does selection for high production affect protein turnover rate? Communication No. 10-03, 7th world congress on genetics applications to livestock production, Montpellier, France

## Books and Reviews

- Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics, 4th edn. Longman Group Ltd., Harlow
- Weller JI (2009) Quantitative Trait Loci Analysis in Animals, 2nd edn. CAB International, Wallingford

## Proton Exchange Membrane Fuel Cells: High-Temperature, Low-Humidity Operation

STEVEN J. HAMROCK<sup>1</sup>, ANDREW M. HERRING<sup>2</sup>

<sup>1</sup>Fuel Cell Components Program, 3M Company, St Paul, MN, USA

<sup>2</sup>Department of Chemical Engineering, Colorado School of Mines, Golden, CO, USA

### Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Proton Exchange Membrane Fuel Cells  
 Electrolyte Membranes  
 PFSA  
 Non-fluorinated or Hydrocarbon PEMs  
 Mechanical Stabilization of Low EW Membranes  
 Stabilizing Low EW Membranes Through Chemical Modification of the Ionomer  
 Conductivity Enhancing/Stabilizing Inorganic Additives  
 Electrodes  
 Future Directions  
 Acknowledgments  
 Bibliography

### Glossary

- 3 Phase boundary** Region in the electrode where protons from the ionomer, electrons from the electrically conducting Pt and/or carbon, and reactant gases meet.
- Electrolyte membrane** A solid polymer ion-conducting membrane used in the center of the fuel cell membrane electrode assembly.
- Fuel cell electrocatalyst** A catalyst that catalyzes either the oxidation of the fuel or the reduction of oxygen in a fuel cell.

**Equivalent weight** A measure of the acid content of an ionomer in the units of grams of polymer per mole of acid.

**Gas diffusion layer** A carbon paper or cloth used as a current collector in fuel cells that can allow the passage of reactant gases and product water to and from the electrodes.

**Hydrogen oxidation reaction (HOR)** Electrochemical oxidation of H<sub>2</sub> at the anode.

**Ionomer** A copolymer of an ion-containing monomer and a nonionic monomer, typically not soluble in water.

**Membrane electrode assembly (MEA)** An ion-conducting membrane sandwiched between two electrodes, an anode at which fuel oxidation occurs and a cathode at which oxygen reduction occurs.

**Oxygen reduction reaction (ORR)** Electrochemical reduction of O<sub>2</sub> at the cathode.

**Flooding** Liquid water collecting within the electrodes or current collectors, impeding the flow of gases to the catalyst surface.

**Perfluorinated sulfonic acid containing polymer (PFSA)** A fluorinated sulfonic acid containing ionomer. The most commonly used polymer in proton exchange fuel cell membranes today.

### Definition of the Subject

Proton exchange membrane fuel cells (PEMFCs) together with hydrogen represent an important storage and utilization technology for energy generated from renewable sources such as wind, solar, geothermal, or hydroelectric. This is due in part to their high energy density, low operating temperature, rapid start-up, modular design, flexibility of scale (a few watts to hundreds of kilowatts), and the absence of any point-of-use emissions. One barrier to commercialization and widespread acceptance of this technology is cost, a situation fairly common with the introduction of a new technology. Over the past decade much work has been done, and very significant progress has been made, in bringing down the manufacturing cost of fuel cell systems [1]. Manufacturing processes have been optimized, volumes manufactured have increased, less expensive materials have been demonstrated, system efficiencies and power outputs have been increased, and the amount of precious metal catalyst required to

generate a kilowatt of power has been reduced dramatically. All of these features have contributed to significant cost reductions. In the case of the precious metal catalysts, one of the major costs, a fuel cell stack that can generate enough power for an automobile can now be built using less than 30 g of platinum catalyst (about 3–4 times as much precious metal as is used in vehicles today), and the auto industry target of 10 g per vehicle appears within reach [2, 3].

There is still work to be done. One area where important improvements are currently being made is in developing materials and constructions that address the need of today's PEMFC systems for high levels of humidification during operation. Materials currently used in PEMFCs require water for optimum performance. The electrolyte membranes require a relatively high level of hydration to provide sufficient conductivity for high performance. The electrodes in use also require water, both to provide ionic conductivity within the electrode and between the electrode and the electrolyte, as well as to maintain high electrocatalytic activity for high efficiency.

This thirst for water within the fuel cell requires strict water management, imposing limitations on the system design and adversely affecting manufacturing cost. Reactant gases entering the cell often must be humidified, adding the expense of humidification equipment and the parasitic power loss from its operation. In addition, cell temperature must be carefully controlled, as overheating can cause the cells to dry out, and so larger cooling systems or radiators are required. This must be balanced with the fact that excess cooling or over humidification can cause water vapor formed in the electrochemical reaction to liquefy and collect within the electrodes or current collectors, impeding the flow of gases to the electrodes, a phenomenon called flooding. These requirements of careful control of humidification and temperature in fuel cells are not consistent with the need for a robust, inexpensive power source. New materials, including new membrane materials and catalysts that are less dependent on water, are needed to address this limitation.

## Introduction

For the last few years, there has been a growing, worldwide public focus on the increasing use of energy. One

cannot pick up a newspaper or watch a television news program without being exposed to stories about the growth in the need for energy, and the economic and environmental cost of that growth. Concerns about energy cost, energy security, and environmental factors (notably climate change) are driving many toward a shift to cleaner, cheaper, and more sustainable methods of generating and using energy. Much of this discussion has centered on the generation of energy, through the more efficient use of fossil fuels, nuclear energy generation, or renewables such as solar and wind energy. There is also growing recognition that if a movement to more sustainable methods of generating energy is to be made, a change in the way of transporting and storing energy will also be required.

An important area of energy technology that has received attention is the area of energy storage. Advanced batteries, capacitors, pumped hydroelectric, compressed air, flywheels, and other methods of energy storage are being considered [4]. As stated above, many believe that proton exchange membrane fuel cells (PEMFCs) together with hydrogen represent an important energy storage and utilization technology for a number of application areas to allow the transition away from fossil fuels. For this reason, significant research and investment in this technology have taken place over the last two decades. PEMFCs are beginning to find use in certain emerging applications, such as backup and primary power supplies for telecommunications, powering material handling fork-trucks and providing electricity in remote, off-grid locations. While these represent relatively low volumes of systems in the greater energy market, they are an enabling first step, which is important for the introduction of fuel cells into a marketplace where they must compete with established technologies.

Another application of PEMFCs that has received much media attention and many research dollars over the past few years is transportation. While fuel cell-powered vehicles are still limited to a few hundred prototypes, they are seen by many as the "end game" for renewable-energy-powered vehicles [3]. Hydrogen fuel cells for powering automobiles are attractive for several reasons. Their high energy density can provide driving ranges of 250 miles or more, and compressed hydrogen tanks can be refilled easily in less than 5 min. This allows automakers to provide vehicles with

essentially the same functionality as drivers enjoy today. However, the strict limitation on weight and volume in automotive applications, as well as the variation in power requirements during use, mean that these systems must run efficiently and reliably under a wide range of temperatures and humidification levels. The current limitations on temperature and humidification require excessively large cooling systems, or radiators, and humidifiers, which make meeting cost and efficiency targets more difficult.

In addition to the utility of hydrogen for storing energy from renewables such as wind or solar, the conversion of hydrocarbon feedstocks such as renewably derived methane or biomass into hydrogen can be an energy-efficient way of utilizing these resources. Currently, the least expensive route to hydrogen is from the reforming of natural gas, a process that initially produces a mixture of hydrogen gas, water, and carbon dioxide with a high carbon monoxide (CO) content. At the current relatively low temperature of operation of the PEM fuel cell, 80°C, CO is a severe poison to the Pt catalyst on the fuel cell anode. It is, therefore, necessary to reduce the CO content of the hydrogen fuel to a few parts per million by use of water gas shift reactors and a final gas clean up stage that may be a hydrogen selective membrane, pressure swing adsorption, or a preferential oxidation reactor. Each of these additional unit operations adds expense to the hydrogen production process. If hydrogen from the reforming of biomass is to be cost competitive, then the tolerance for CO on the catalyst must be improved so that less expensive less pure hydrogen can be utilized. One method of doing this is to operate the fuel cell at elevated temperature, e.g., a phosphoric acid fuel cell operating at >180°C can tolerate a reformed hydrogen fuel containing 2% of CO. The operation of PEM fuel cells at elevated temperatures would, therefore, enable the utilization of biomass-derived hydrogen at a price competitive fuel cost.

For the rate of commercialization of PEMFCs to continue to increase, system costs must continue to decrease. One way to do this is to eliminate the fuel cell temperature and humidification requirements described above, allowing operation over a wide range of temperatures without the need for humidification of the incoming gases. To do this, new materials are needed. These include new ion-conducting materials

for membranes and electrodes and new catalysts that can function with less water. This entry will review how these materials function in a PEMFC and some of the approaches to new materials that may overcome the humidification and temperature barriers.

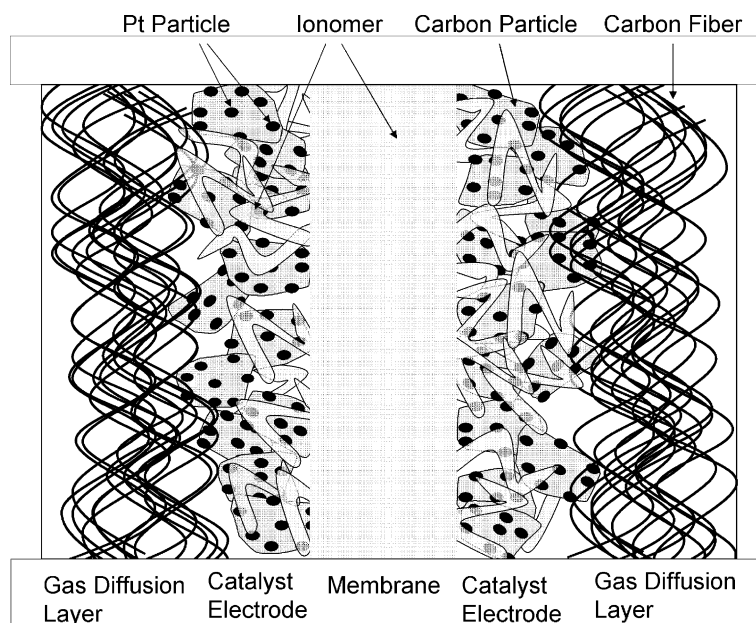
## Proton Exchange Membrane Fuel Cells

A fuel cell is an electrochemical cell that oxidizes a fuel to provide electrical energy. It is similar to an engine in that you provide fuel and air to generate energy, and it is similar to a battery in that it is an electrochemical cell that produces electricity. A variety of fuels can be used depending on the type of fuel cell. High temperature fuel cells, such as solid oxide fuel cells, can use a wider assortment of fuels because the electrocatalysts are more efficient and less prone to poisoning at these higher temperatures, up to 1,000°C. These fuels include hydrogen, alcohols, and hydrocarbons. Fuel cells that operate at lower temperatures are typically restricted to using fuels that are more easily oxidized, such as hydrogen or methanol. PEMFCs fall into this class. PEMFCs use a polymeric ion exchange membrane as an electrolyte and operate at lower temperatures, typically up to about 80°C.

A schematic of the cross section of a single cell, often called a *membrane electrode assembly* (MEA), is shown in Fig. 1. The electrolyte membrane is at the center of two porous, catalyst-containing electrodes. The electrodes are typically formed from carbon-supported platinum particles. These carbon particles are held together by a small amount of an ion-conducting polymer, which act as both a binder and an ion conductor, allowing protons to move through the electrode. Newer types of electrode structures that allow for improved catalyst efficiency and durability are being introduced [5]. This three-layer construction is then positioned between two porous gas diffusion layers that act as current collectors. Hydrogen is supplied to the negative electrode, or anode, and oxygen, usually in the form of air, is supplied to the positive electrode, or cathode. The product water is formed at the cathode.

A PEMFC system typically comprises a fuel cell stack where MEAs are stacked between electrical conductive bipolar plates that have flow fields embedded in them, allowing the reactive gases to be supplied to the catalyst surface and allowing the reactant water to be





**Proton Exchange Membrane Fuel Cells: High-Temperature, Low-Humidity Operation. Figure 1**  
A schematic representation of the cross section of an MEA

carried away. The area of the MEA determines the amount of current that can be passed through a cell and the number of cells in the stack determines the voltage. Together, these define the power the stack is capable of providing.

To allow PEM fuel cells to operate under the hotter, drier conditions required for widespread use in applications such as automobiles, new materials are needed. These include new electrolytes with higher proton conductivity and improved durability at low relative humidity (RH) and at higher temperatures. New electrodes that can provide adequate performance with less water are needed.

### Electrolyte Membranes

The electrolyte in a PEMFC, as the name implies, is a proton exchange membrane, or PEM. It functions by allowing transport of protons from the negative to positive electrode, and as a physical barrier to prevent shorting of the electrodes and crossover of the reactant gases. The requirements for an electrolyte membrane in a PEMFC typically include the following:

- High proton conductivity
- Low permeability to reactant gases

- Good mechanical properties both dry and equilibrated with water
- Stability toward leaching of components by liquid water
- Excellent chemical stability (Hydrolytic and oxidative)
- Reasonable cost
- The ability to form stable intimate interfaces with the electrodes

A variety of types of materials have been used in electrolyte membranes for PEM fuel cells. Most of these fall into two classes: basic polymers that have been imbibed with an acid and polymers with acidic groups attached.

In the first category, the most commonly used polymers for this are polybenzimidazole (PBI) or analogs imbibed with phosphoric acid [6]. This type of membrane was developed at Case Western Reserve University in the mid-1990s [7]. These membranes are known to have good conductivity at very high temperatures, up to 200°C, and membranes with high phosphoric acid contents, and increased conductivity combined with good mechanical properties have been prepared [8]. MEAs comprising such membranes are commercially available from the BASF company.

There are drawbacks to using PBI/Phosphoric acid-based membranes in many fuel cell applications. The highly water-soluble phosphoric acid can be easily leached out of the membrane by liquid water, preventing use in applications where the cell could experience higher humidifications or lower temperatures. The phosphoric acid also adsorbs to the platinum catalyst surface, inhibiting the electrode kinetics, particularly on the oxygen electrode. To overcome this, high levels of expensive platinum catalysts are required for adequate fuel cell performance. It should be stressed that while PBI/Phosphoric acid-based membranes have drawbacks that prevent their widespread use, they are the only commercially available membranes that can be used in the temperature range between about 120°C and 200°C.

The majority of PEMs used today are from the second class of polymers, those with pendant acidic groups. Specifically, most polymers currently used in PEMs are typically members of a class of polymers called *ionomers*. An ionomer is a copolymer of a strong acid containing monomer and a nonionic, neutral monomer [9]. When the neutral monomer is relatively nonpolar, ionomers will adopt a phase-separated morphology, where the ionic groups can bind tightly together into ionic aggregates or clusters. These clusters have a significant impact on the physical properties of the ionomer, often behaving as physical cross-links and stiffening the polymer [10]. A few examples of ionomers of this type are commercially available, such as DuPont's Surlyn™, a copolymer of ethylene and a salt of methacrylic acid, which is used in several applications, including the coating on the outside of golf balls.

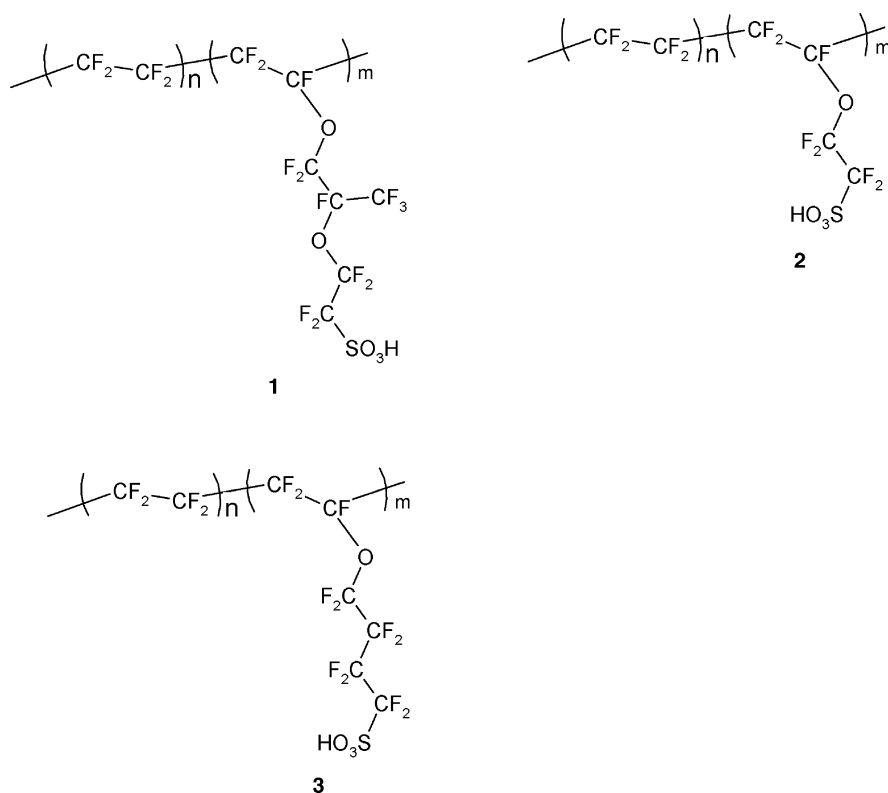
In order for the protonated form of an ionomer to be suitable for use in a fuel cell, it must be chemically and mechanically stable enough to survive the chemically aggressive, oxidizing environment of a fuel cell. Oxidizing species such as peroxides can be formed during operation, which attack and chemically degrade the membrane [11]. Simultaneously, the membrane is mechanically stressed from the fluctuations in water content resulting from variations in current density and temperature. These combined can cause the membrane to fail, leading to gases crossing over and catastrophic cell failure. For this reason, ionomers used in fuel cells today fall into two categories of polymers that

have sufficient chemical stability and mechanical properties. These are perfluorinated sulfonic acid containing polymers (PFSA) and aromatic-backbone polymers with pendant sulfonic acid groups.

### PFSA

Perfluorinated sulfonic acid containing polymers (PFSA) are the most commonly used membrane materials in fuel cells today. Membranes made from these ionomers provide the benefits of highly acidic pendant acid groups for high proton conductivity, good mechanical properties, excellent chemical stability, and fairly low cost. The first PFSA used in PEMFCs was DuPont's Nafion™, originally developed in the 1960s for brine electrolysis to produce chlorine [12]. Since then, several other PFSA membranes have been developed and introduced for use in fuel cells [13]. All of these are copolymers of tetrafluoroethylene (TFE) and a sulfonic acid containing monomer. The chemical structures of some of these polymers are shown in Fig. 2.

When enough acid groups are present in the ionomer, the very hydrophilic sulfonic acid aggregates will absorb water. These hydrated acid groups can provide a continuous, acid-rich, hydrated pathway through the polymer. For PFSA, in addition to the ionic regions, the TFE segments in the backbone provide another structural feature of the polymer. If the ratio of TFE units to acid containing monomers is high enough to provide TFE runs of sufficient length (about 4 or more TFE monomer units) these can crystallize, much like the highly crystalline polymer, polytetrafluoroethylene. These crystallites in the hydrophobic region of the polymer provide significant mechanical stabilization to the membrane. The amount of acid contained in the membrane is typically expressed as *equivalent weight* (EW), the number of grams of polymer required to provide one mole of acidic protons. For traditional PFSA such as Nafion™, this value is in the range of 1,000–1,100. This gives a ratio of TFE to acidic monomers of about 5.5–6.5, enough to provide some stabilizing backbone crystallinity. Such polymers have a good combination of proton conductivity and mechanical properties when fully hydrated. Much work has been done over the years to provide a detailed understanding of the structure of PFSA, and this is



### Proton Exchange Membrane Fuel Cells: High-Temperature, Low-Humidity Operation. Figure 2

Structures of some perfluorinated sulfonic acid containing polymers (PFSA). Polymer 1 is available from DuPont (Nafion™), Asahi Glass (Flemion™), and others; Polymer 2 is the short-side-chain ionomer developed at Dow, currently available from Solvacore; and Polymer 3 is the ionomer available from 3M Company

still an active area of research. A comprehensive review on the subject has been written by Mauritz and Moore [14]. A generalized representation of a hydrated PFSA structure is shown in Fig. 3.

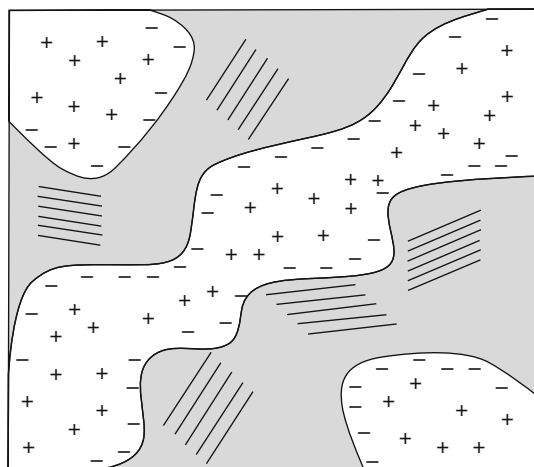
The amount of water present in the hydrated channels of the membrane is a function of the number of sulfonic acid groups present in the membrane and the humidity of the reactant gases [15]. In a typical PFSA membrane, at a given relative humidity the ratio of water molecules to sulfonic acid groups (referred to as lambda,  $\lambda$ ) is fixed. At low %RH, there are a few tightly bound water molecules. As the %RH is increased, more water is absorbed and these additional water molecules are less tightly bound and more mobile. It is thought that the less tightly bound, more mobile water molecules that are farther from the sulfonic acid groups are more able to contribute to proton transport [16]. When the temperature is increased, or humidity levels in the reactant gases are decreased, the membrane will

dry out and the conductivity drops. This represents an increase in the resistance of the cell and causes a loss in efficiency and performance.

One method of maintaining high conductivity with less water is to lower the EW, increasing the concentration of sulfonic acid groups in the membrane.

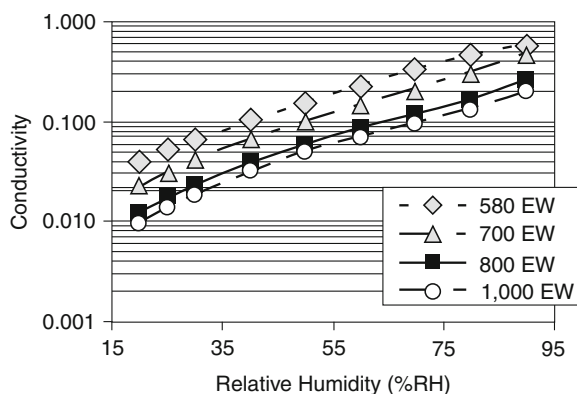
Figure 4 shows the conductivity as a function of relative humidity for several different EW membranes. Lower EW membranes do provide a significant increase although conductivity still drops off at lower relative humidity.

Another way to consider the impact of membrane conductivity on fuel cell performance is shown in Fig. 5. Figure 5a shows the conductivity of a few different EW membranes as a function of temperature with the atmosphere inside the conductivity cell held at a fixed dew point of 80°C [17]. When the conductivity cell is at 80°C, the %RH is 100%. As the temperature of the cell increases, the %RH at a fixed dew point



**Proton Exchange Membrane Fuel Cells: High-Temperature, Low-Humidity Operation. Figure 3**

A general representation of the morphology of a hydrated PFSA. The + represents the hydrated protons and the – represents the sulfonate groups at the edges of the hydrated region. These hydrated regions are thought to have the dimensions of 2–3 nm. The parallel lines represent the crystallites formed from the tetrafluoroethylene (TFE) groups of the backbone



**Proton Exchange Membrane Fuel Cells: High-Temperature, Low-Humidity Operation. Figure 4**

Conductivity vs. relative humidity for several different 3M ionomers (Polymer 3) measured by AC impedance spectroscopy at 80°C

decreases, causing a decrease in the membrane conductivity. This is similar to the situation in some PEMFC applications where the cell temperature may rise while the humidity level of the incoming gases remains

constant. The graph in Fig. 5b uses the same data. Here the conductivity is used to calculate the resistance of a 25  $\mu\text{m}$  membrane, and using Ohm's law, that resistance is used to calculate the voltage loss (ohmic loss) one would see in a fuel cell at a 0.6  $\text{A}/\text{cm}^2$  current density [17]. This represents the fuel cell performance loss due to the loss of membrane conductivity (certainly not the only performance loss under these conditions!).

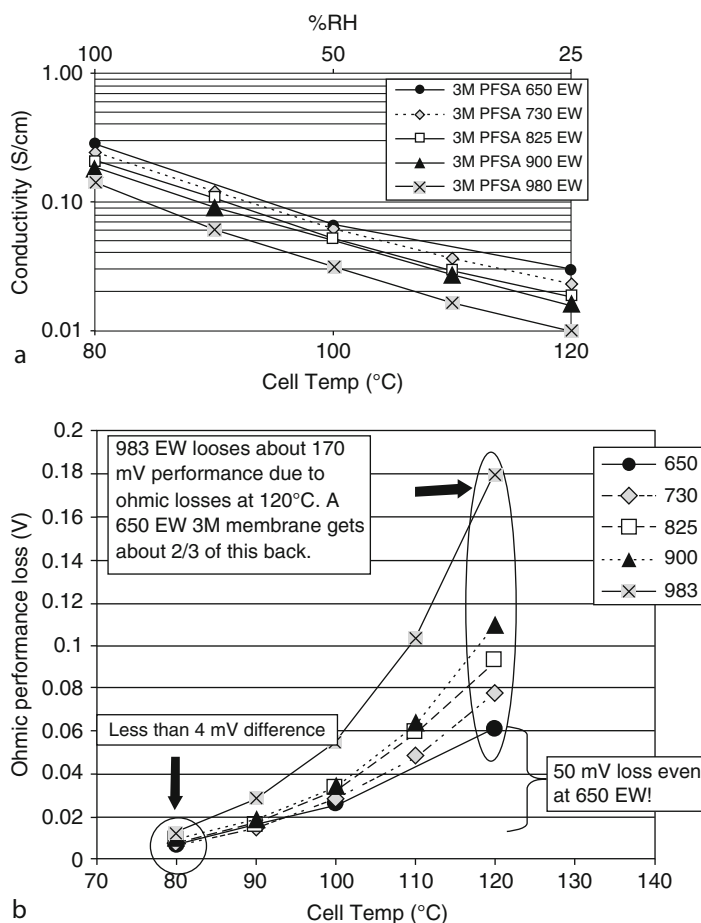
At 80°C, 100% RH, the performance loss is low, about 10 mV. Further, the performance difference between the different EW membranes is also quite low, less than 4 mV. As the temperature increases, the performance losses also increase and the effect of the different EW ionomers becomes apparent. At 120°C, the 1,000 EW membrane has a large ohmic loss of about 180 mV. This represents a  $\geq 20\%$  loss in the operating voltage of a typical PEMFC at this current density, or about  $\geq 15\%$  of the energy contained in the hydrogen fuel being converted to heat. The lower EW membranes do provide a significant improvement, but even at the lowest EW shown here, 650, the ohmic loss is still 6 times that of the fully humidified cell.

Lowering the EW of the ionomer does seem to provide at least a partial solution to this problem. This suggests the possibility that even lower EW ionomers could allow performance equivalent to the fully hydrated membranes even under these dry conditions. In the case of typical PFSA's, this is not a practical approach. This is due in part to the lack of the backbone crystallites mentioned above. At an EW of below about 700, these polymers do not have enough TFE to provide sufficient backbone crystallinity. This renders the membrane effectively water soluble, and thus not useful in most PEMFC applications [18]. The solubility of the 3M ionomer as a function of EW is shown in Fig. 6.

PFSA ionomers with lower MW side chains should allow additional crystallinity at a given EW. The Polymer 2, shown in Fig. 2, has a side chain that is 100 MW units lower than the 3M ionomer, so it may allow a more stable membrane at somewhat lower EW.

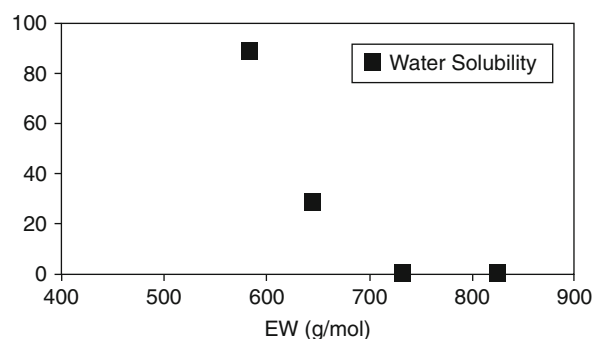
### Non-fluorinated or Hydrocarbon PEMs

A variety of non-fluorinated or partially fluorinated ionomers have been evaluated as alternatives to PFSA's for PEM fuel cells. These are typically sulfonated



**Proton Exchange Membrane Fuel Cells: High-Temperature, Low-Humidity Operation. Figure 5**

(a) Conductivity of various equivalent weight (EW) 3M ionomer membranes (Polymer 3) as a function of temperature in an atmosphere with an 80°C dew point. (b) Calculated performance loss due to membrane ohmic losses at 0.6 A/cm<sup>2</sup> for these membranes at 25 μm thickness



**Proton Exchange Membrane Fuel Cells: High-Temperature, Low-Humidity Operation. Figure 6**

Water solubility of 3M ionomers (Polymer 3) as a function of EW. The samples were boiled for 3 h

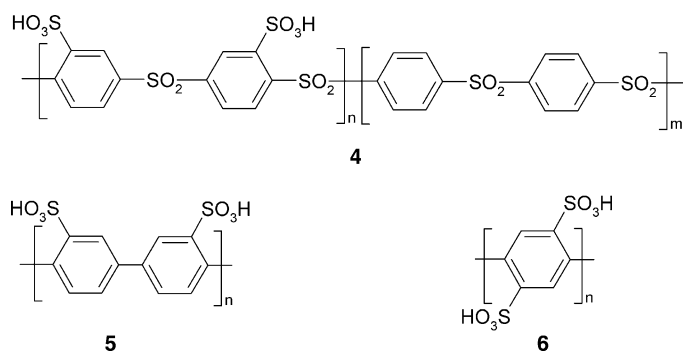
aromatic hydrocarbon polymers. Examples include sulfonated engineering thermoplastics such as polyimides [19], polyetherketones [20], and polysulfones [21] as well as polyphosphazines [22], or sulfonated polystyrene grafted to fluoroplastics such as polyvinylidene fluoride [23]. Some of the observed or proposed advantages of these membrane materials include lower cost, increased toughness or improved mechanical properties, and lower permeability to oxygen and fuels [24]. Permeation of oxygen through the membrane is thought to lead to formation of hydrogen peroxide on the hydrogen electrode, contributing to chemical degradation of the membrane [25]. One significant advantage of hydrocarbon-based ionomers

that PFSA's do not have, is their inherent synthetic versatility, allowing one to more easily design the polymer structure one needs for optimum conductivity, physical and mechanical properties, and chemical stability (of course, this assumes one knows what structure one needs).

Many examples of hydrocarbon ionomers have been prepared by exposing aromatic-backbone polymers to sulfonating agents, producing ionomers with sulfonic acid moieties attached to the most electron-rich positions on the aromatic rings. By controlling the degree of sulfonation, ionomers can be prepared by this method with suitable swelling characteristics and high proton conductivity at high relative humidity. It should be pointed out that due to the lower density of hydrocarbon-based polymers compared to fluoropolymers; a lower EW is needed in a hydrocarbon ionomer to provide an equivalent volumetric density of acid groups of a PFSA. Differences in the volumetric density of acid groups are more useful when comparing the conductivity of ionomers based on different classes of polymers [26]. Unfortunately, many studies have shown that randomly sulfonated hydrocarbon ionomers often suffer from lower conductivity at low relative humidity compared to PFSA's [20]. This is likely a consequence of a less favorable microstructure for proton transport as well as the lower acidity of the sulfonic acid groups bound to the aromatic ring ( $pK_a = ca. -2$  to  $-4$ ) compared to the sulfonic acid groups of the PFSA ( $pK_a = ca. -5.5$ ) [20, 27].

Synthetic methods that allow attachment of sulfonic acid groups to more electron-deficient sites on

aromatic rings can produce polymers where these groups are not only more acidic, but also more stable toward thermal desulfonation [28]. Kreuer and coworkers have prepared such polymers in sulfonated polysulfones with a variety of EWs [29]. The structure is shown as Polymer 4 in Fig. 7. These ionomers have high thermal stability, high conductivity at low levels of hydration, and surprisingly low water solubility. These sulfonated polysulfones are not water soluble at  $100^\circ\text{C}$  at EW values of down to 430. The improved conductivity at low hydration is likely due in part to the low electron density of the aromatic rings of the polysulfone, and also possibly due in part to a favorable microstructure for proton transport. To further increase conductivity under very dry conditions, some highly sulfonated hydrocarbon polymers have been shown to have very high proton conductivity, even at low relative humidity. The Kreuer group has also prepared a completely sulfonated polysulfone, that is polysulfone with a sulfonic acid group on every aromatic ring (EW = 220). While this ionomer is water soluble, it has conductivity substantially higher than 1,100 EW Nafion™, even under dry conditions [30]. Litt and coworkers have prepared highly sulfonated polyphenylenes with one and two sulfonic acid groups per aromatic ring, shown as polymers 5 and 6 in Fig. 7 [31]. The latter has an EW of 118! Both of these ionomers are also water soluble and have very high proton conductivity at low %RH. These ionomers have been shown to hold more water at lower %RH than other sulfonic acid-based ionomers. This observation was explained by an increase in the “frozen-in



**Proton Exchange Membrane Fuel Cells: High-Temperature, Low-Humidity Operation. Figure 7**

Structures of some hydrocarbon ionomers. Polymer 4 is from reference 30 and Polymers 5 and 6 are from reference [32]

free volume” in these ionomers, that is that the rod-like morphology of the polymer hindered close packing. Removal of the last few waters of hydration in the voids between these rods would force them closer together into a higher energy state, effectively increasing the heat of vaporization of the bound water molecules.

The studies mentioned above show that through control of the electronic and structural features of hydrocarbon ionomers, increased conductivity can be achieved. However, this is often at the expense of the mechanical stability of the polymer to the point where these materials cannot be used in fuel cells. One potential method of stabilizing these materials is to incorporate them into a stable multiphase or segmented system. A variety of synthetic methods exist that allow generation of different branched or block copolymers and these have been applied to the synthesis of PEMs [32]. This allows control over the morphology of the phase-separated structures to create interconnected proton-conducting channels that may allow increased proton conductivity. McGrath and coworkers have prepared and evaluated sulfonated multiblock polyarylene ether sulfones with conductivity at low relative humidity equivalent to a Nafion™ membrane [33, 34]. The conductivity has been shown to be a function of the length and the chemistry of both the hydrophilic and the hydrophobic blocks.

### Mechanical Stabilization of Low EW Membranes

One way of mechanically stabilizing low EW ionomer membranes is to generate a composite membrane using a porous film as an internal reinforcing structure [35, 36]. A PFSA membrane reinforced with a thin expanded polytetrafluoroethylene layer is available from W.L. Gore. Reinforced membranes of this type have been shown to have increased strength and lower in-plane swelling upon hydration, lowering the potential of damage due to stresses generated during fuel cell operation. This should result in increased fuel cell durability [37]. Composite membranes have also been formed by using the porous phase as the conducting phase and filling the pores with a reinforcing phase. Pintauro and coworkers have made membranes using microfibers of sulfonated polyether sulfone filled with an inert filler to provide a membrane with good

mechanical properties and proton conductivity when fully hydrated [38]. This group then used low EW PFSA fibers in this process, which gave a membrane with low swelling and very good conductivity at relatively low % RH (0.10 S/cm conductivity at 80°C and 50% RH, about 2–3 times higher than a 1,100 EW Nafion™ membrane) [39].

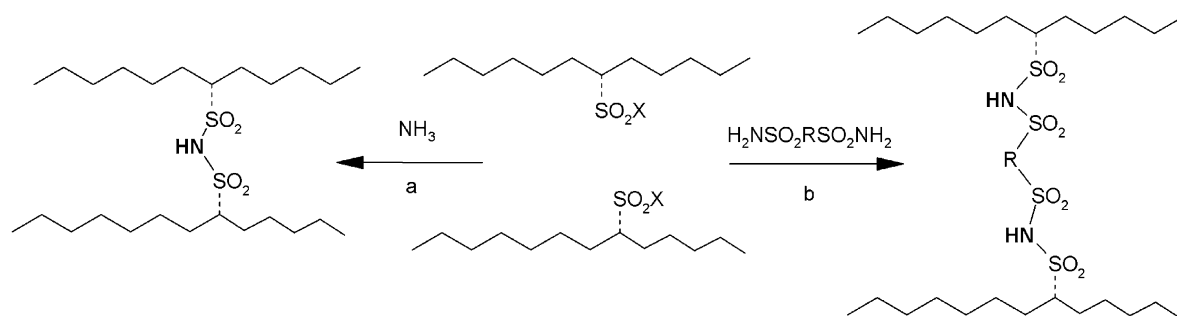
Mechanical stabilizing membranes in this way can allow significant stabilization of low EW ionomers. However, since many applications of PEM fuel cells require not only hotter and/or drier operation conditions but also require that the membrane to be insoluble in liquid water (often hot!) at times during operation, there is a limit to how low this method will allow one to go. In order to allow very low EW ionomers to be feasible, a change in the polymer chemistry will also probably be required.

### Stabilizing Low EW Membranes Through Chemical Modification of the Ionomer

One possible method of chemically stabilizing low EW ionomers's toward excessive swelling and dissolution in water is to cross-link the ionomer. Some attempts are being made to cross-link low EW ionomers. [40, 41]. Generally, there are two “regions” in which ionomers can be cross-linked, in the hydrophilic, conducting region, near the acid groups and in the hydrophobic region, near the backbone. In the case of the former, one method that has been studied is forming a bis-sulphonyl imide from two of the pendent sulfonyl halide groups on the ionomer precursor [42]. Bis-sulphonyl imides are known to have highly acidic protogenic hydrogens and excellent chemical stability [43]. This method has the advantage that the cross-links formed have similar acidity to the acid groups consumed. A generalized representation of this method is shown in Fig. 8.

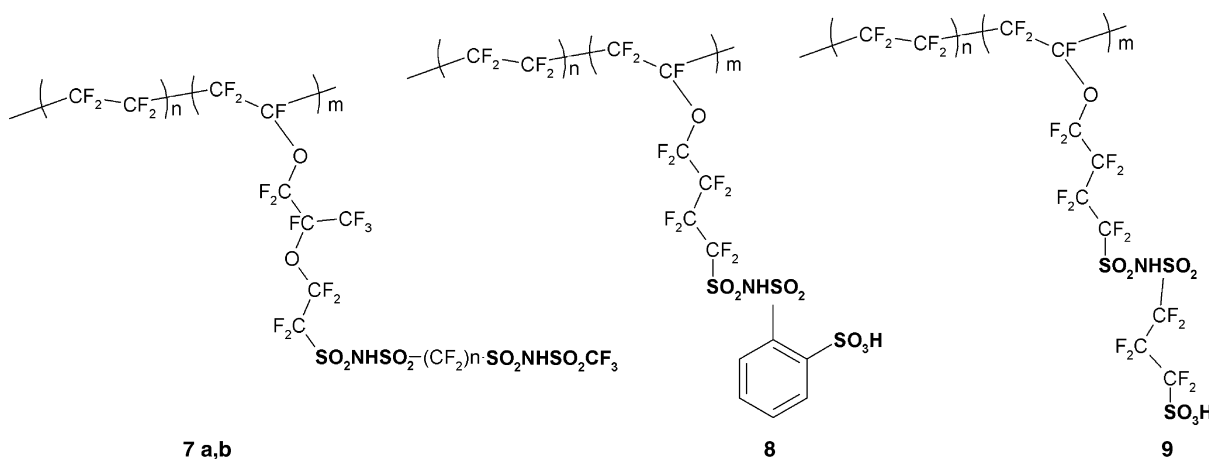
Methods in which the backbone of the polymer can be cross-linked include radiation grafting [44] and through the preparation of a cross-linkable terpolymer by including a reactive third monomer in the polymerization of the ionomer, followed by curing in film form [45–47].

Another approach to providing ionomers with lower EW and suitable mechanical and solubility properties is to have more than one acidic proton per side



Proton Exchange Membrane Fuel Cells: High-Temperature, Low-Humidity Operation. Figure 8

General method of synthesizing bis-sulfonyl imide containing cross-links from polymers with pendent sulfonyl halide groups



Proton Exchange Membrane Fuel Cells: High-Temperature, Low-Humidity Operation. Figure 9

The structure of 3 multi acid side-chain ionomers. Polymer 7a,  $n = 4$ , 7b  $n = 2$  are from reference 49, Polymers 8 and 9 are from reference [50]

chain. If the side chain has additional protogenic groups, a low EW ionomer can be prepared having a higher degree of backbone crystallinity, and hopefully increased stability toward liquid water. One way to prepare such ionomers is to include a highly acidic bis-sulfonyl imide in the side chain. Such materials were prepared by Desmarteau (Polymers 7a,b) and more recently at 3M (Polymers 8 and 9) [48, 49, 51]. The structures of some of these materials are shown in Fig. 9.

The relationship between the number of TFE units that form the backbone crystallites and EW is shown in Fig. 10. The slope of each line gives the EW of the ionomer/the ratio of TFE units to protons in the polymer, and the intercept is the MW of the acid functional

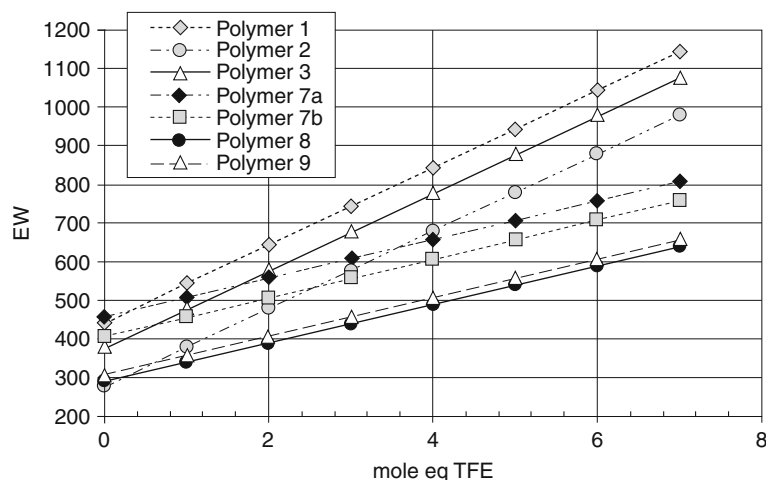
monomer/the number of protons. This shows the utility of having multiple protogenic groups on each side chain in providing polymers having high crystallinity and low EW.

In the case of Polymers 8 and 9, it has been demonstrated that low EW ionomers with higher conductivity, low swelling in boiling water, and good mechanical properties can be prepared [50].

### Conductivity Enhancing/Stabilizing Inorganic Additives

Another approach to overcome the inherent deficiencies of ionomers under hot, dry operating conditions





Proton Exchange Membrane Fuel Cells: High-Temperature, Low-Humidity Operation. Figure 10

Plot of EW vs. the number of TFE units in the backbone of selected ionomers

has been to investigate the use of inorganic additives to form composite membranes [51]. Three basic functionalities are invoked:

1. Additives that are hygroscopic and designed to retain additional water in the membrane so that no loss of performance is observed when the fuel cell is operated under conditions of reduced RH. If the fuel cell spends significant time under dry operations, these approaches inevitably fail.
2. Additives that have enhanced acidity and can facilitate proton transport and so enhance performance under drier conditions.
3. Additives that are designed to decompose peroxide in situ in the membrane to increase the membrane chemical durability.

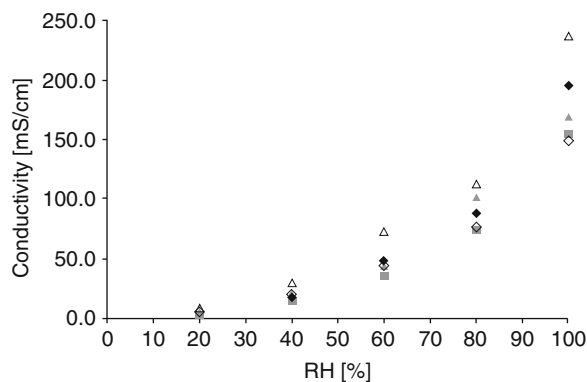
A number of additives also have combined functionality.

Probably every common hygroscopic inorganic oxide has at some time been used to prepare a composite membrane for fuel cell use. The perceived benefit of an insoluble inorganic additive is from a surface interaction between the additive particle and the ionomer and so nano-sized individual particles or meso-porous materials into which the ionomer can penetrate have the larger benefit. Larger particles give no additional benefit and simply reduce the EW of the ionomer. The inorganic materials are either preformed before being mixed with the ionomer or are formed in

situ, typically by a solgel process. Of the oxides that can be formed in situ, the most commonly used additive has been silica, but it is unstable to acid and so its suitability for fuel cell operation is questionable [52]. Titania and zirconia composites would appear to have more promise from a stability viewpoint [53], although they have mostly been found to enhance membrane mechanical properties, as ultimately the water in these additives will also be lost on sustained dry operation. Recently improved performance has been observed under drier operation by combining tin oxide with titania [54]. Clays both natural and synthetic have also been used, but again their benefit to fuel cell under RH cycling is also questionable.

More promising are approaches using either acid-functionalized particles [55] or super acidic inorganic materials that are designed to increase proton mobility. Of these, the two most promising are zirconium phosphonates and the heteropoly acids (HPAs) [56–58]. The effect of these may simply that they are more hygroscopic, or that the phenomena is simply a proton concentration effect, essentially lowering the equivalent weight. However, as they also lower the activation energy for proton transport it seems that they also act as an effective proton transport promoter, perhaps more effective than the sulfonic acids.

Figure 11 shows proton conductivity data at 100°C for the 3M ionomer doped with various HPAs. Two observations are immediately apparent: (1) that the



**Proton Exchange Membrane Fuel Cells: High-Temperature, Low-Humidity Operation. Figure 11**

Proton conductivity at 100°C for the 3M ionomer (■) and the 3M ionomer doped with (◇) 1% HPW, (◆) 5% HPW, (△) 1% HSiW, and (▲) 5% HSiW at 20–100% RH. From reference [59]

structure and amount of the additive have a strong influence and (2) that the effect becomes dramatically less as the RH is lowered. Similar results are shown for zirconium phosphonate composites with PFSA ionomer [59].

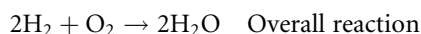
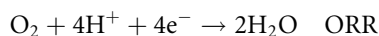
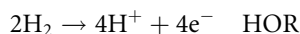
Hydrogen peroxide decomposition catalysts can be added to ionomer membranes in small amounts to slow down the decomposition of the ionomer during fuel cell operation. Additions of cerium and manganese, in both oxide and ionic forms, have been shown to increase the oxidative stability of membranes by orders of magnitude, and fuel cells prepared with such membranes have shown substantial increases in lifetime under aggressive hot and dry operation [60–62]. Unfortunately, these metal ions and oxides can consume ion exchange capacity and negatively impact fuel cell performance.

The ideal additive would enhance proton conductivity and stability. One demonstration of this was in a composite PFSA membrane using Pt nanoparticles supported on titania or silica [63]. The composite membranes when employed in MEAs demonstrated unhumidified fuel cell performance comparable to that of a similar humidified fuel cell. Whether adding Pt to the membrane will help durability or hurt it is still a matter of some debate [64, 65]. Unfortunately, it is not commercially feasible at this time to add additional Pt to the MEA and so this approach while novel is not

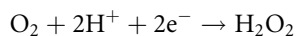
practical. The HPAs are known peroxide decomposition catalysts and so these inorganic oxides have been demonstrated to improve performance and decompose peroxide in fuel cells and if they could be immobilized would present a practical solution to this problem [66].

## Electrodes

As stated above, the membrane acts as the proton transporting medium, is an electrical insulator, and separates the reactant gases from direct chemical reaction. On either side of this membrane are placed two electrodes. The anode at which hydrogen is consumed in the hydrogen oxidation reaction (HOR) and the cathode in which oxygen from air is consumed in the oxygen reduction reaction (ORR). The two half-cell reactions and the overall reaction are shown below.



The electrons flow around an external circuit and do work, while the protons pass through the fuel cell membrane. This overall reaction represents the combustion of hydrogen that produces heat, one reason why the membranes function to separate the two reactant gases is critical. Not only would a leak lead to fuel cell inefficiency, but also a hot spot would develop at the site of the leak, which would result in potential damage to the MEA. Unfortunately, the ORR is not 100% efficient and a  $2\text{e}^-$ ,  $2\text{H}^+$  reaction results in the formation of hydrogen peroxide as shown below:



This reaction is currently unavoidable and appears to be favored at hot and dry operating conditions of the fuel cell. The peroxide decomposition forms reactive radicals such as hydroxyl,  $\bullet\text{OH}$ , and peroxy,  $\bullet\text{OOH}$ , that cause oxidative degradation of both the fuel cell membrane and catalyst support [67]. Both electrodes currently use Pt or Pt alloys to catalyze both the HOR and ORR reactions. The catalyst particles are typically supported on a high surface area, heat-treated carbon to both increase the effectiveness of the catalyst and to provide a path for the electrons to pass through to the external circuit via the gas diffusion media (which is

typically also made of carbon) and the current collecting bipolar plates. In addition, the catalyst particles are coated in ionomer to facilitate proton transport; however, the electrode structure must also be porous to facilitate reactant gas transport. A schematic of a typical PEM MEA is shown in Fig. 1. A boundary condition exists at the catalyst particle where protons from the ionomer, electrons from the electrically conducting Pt and carbon, and reactant gases meet. This is usually referred to as the *three-phase boundary*. The transport of reactants, electrons, and protons must be carefully balanced in terms of the properties, volume, and distribution of each media in order to optimize operation of the fuel cell.

Typically, a good proton conductor is thought to be one where the proton conductivity is  $\leq 0.1 \text{ S cm}^{-1}$ , however, from the point of view of fuel cell operation it is the area-specific resistance (ASR) of the MEA that is more important. If one was to consider the MEA as a series of resistances an anode resistance would be observed, an interfacial resistance between the membrane and the anode, a membrane resistance, an interfacial resistance between the membrane and the cathode, and a cathode resistance. It is assumed here that the resistance of electrical connection between the anode and the current collectors is negligible compared to those described above, however, this too can be compromised if there is insufficient pressure between the bipolar plates and the gas diffusion media. All of these resistances must be optimized in order to lower the area-specific resistance of the fuel cell. The effect of the resistances, or ohmic losses on the overall performance and efficiency of the fuel cell is illustrated in Fig. 5. While a large amount of current work is concerned with optimizing membrane ionomers for hotter and drier operation, little thought has to date been put into optimizing the electrode ionomer, the ionomer catalyst interface, or the catalytic reactions at the anode or cathode for higher temperature, lower RH operation. If the ionomer in the membrane is not well matched and linked to that in the fuel cell electrodes, a large ASR can result. Of course part of the reason for this, until recently, has been the lack of suitable hot, dry ionomers for practical fuel cell testing.

As stated above, in a conventional, fully humidified fuel cell, part of the reactant gas stream is diluted by water vapor and the cathode suffers from formation of

liquid water blocking the pores, or flooding, as the water is being produced in a water-saturated environment. To overcome this problem, hydrophobic fillers such as Teflon™ may be added to the electrode to facilitate water rejection [68]. These systems have been to a large extent already optimized, and great deal of art pertains to electrode fabrication [69]. One advantage of running a fuel cell hot and dry is that the electrode flooding issue is eliminated. In these fuel cells, there is still water produced on the cathode, but possibly not enough to saturate the PFSA polymer in the electrode layer, and potentially leaving the anode side of the fuel cell under humidified. However, as it is likely that polymers with low EW will be used for high temperature operation, back diffusion of water should be increased improving the chance that the anode will not be dried out. With less water in the fuel cell system, freeze issues on start-up in cold climates may also be partially mitigated.

Each of the electrode components is now considered in terms of hot and dry operation, what is known, and what needs to be accomplished to realize these systems.

It is generally thought that as temperature increases so do reaction kinetics. However, the situation in a fuel cell, an electrochemical device, is far more complicated. The reaction mechanism will depend on the surface environment of the catalyst particle and the potential at which the reaction is taking place. The electrode overpotential associated with the ORR represents the largest voltage loss in fully humidified fuel cells and so it is important that the situation not be exacerbated in running fuel cell under hot and dry conditions. In the kinetically controlled region of the fuel cell operation, the performance can be described by the Tafel equation:

$$E = E_{\text{rev}} + b \log i_0 - b \log i$$

$$b = -2.3 \frac{RT}{\alpha n F}$$

where  $E$ ,  $E_{\text{rev}}$ ,  $b$ ,  $i$ ,  $i_0$ ,  $n$ , and  $\alpha$  are the electrode potential, reversible potential, Tafel slope, current density, exchange current density, the number of electrons transferred in the rate determining step, and the transfer coefficient, respectively [70, 71]. The first observation is that increasing the fuel cell temperature from 60 to 120°C, while maintaining a constant relative

humidity (RH), causes the theoretical open circuit voltage (OCV) to decrease from 1.22 to 1.14 V due to the increase in water partial pressure [72] and so again it is desirable to operate the fuel cell at reduced RH. The Tafel slope, a measure of the potential loss of the electrode due to reaction kinetics, is the logarithmic decrease in current density with applied voltage. It is therefore desirable to have as small a Tafel slope as possible. The Tafel slope varies with current density as the surface of the platinum varies with voltage. At high voltage or low current density, the Pt is oxide coated (Temkin adsorption conditions) and the Tafel slope is 60 mV/decade; at lower voltage, higher current density the Pt is oxide free (Langmuir adsorption conditions) and the Tafel slope is 120 mV/decade. So above 100°C, the reaction mechanism may change if the surface coverage is compromised by the lack of water. Experimentally it has been shown for a water-saturated electrode that the Tafel slope increases with temperature at high voltage but is invariant at low voltage [73].

There are very few studies of the ORR under hot and dry fuel cell operating conditions. Recently, methods have been devised to separate the mass transport effects from the kinetic effects [74, 75], but none of these have been applied to hot and dry fuel cell operation. These studies showed that, under fully humidified conditions up to 70°C, oxygen reduction had a tenfold higher specific performance for platinum black at 0.90 V compared to Pt on carbon as has been previously reported in the literature [76]. However, this significant benefit of platinum black is shown to rapidly decrease when the potential is shifted to lower, more fuel cell relevant potentials. This is manifested in the Tafel slope, which decreased from ~360 to ~47 mV/decade in the region where the overpotential was <0.35 V. The effect of hot and dry conditions has been studied in a 5 cm<sup>2</sup> MEA where mass transport and kinetics are difficult to separate [73]. At 120°C, the Tafel slope is found to increase inversely with RH. It is speculated that this is due to the decrease in ionic conductivity in the electrode. RH can also influence water oxidation to form Pt-OH and Pt-O and thereby change the surface condition of the platinum crystals.

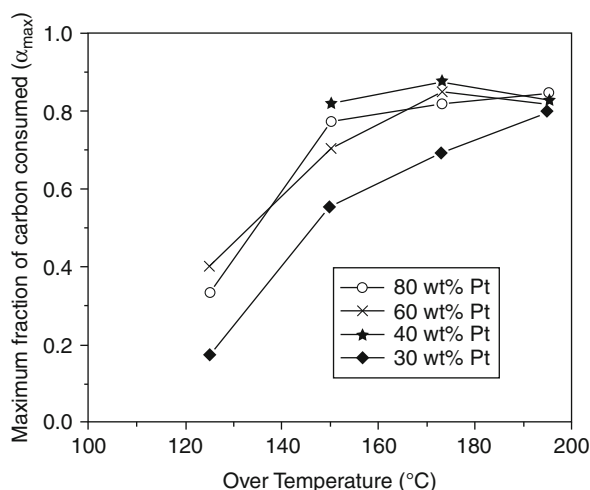
It has been shown that the current exchange density increases up to 70°C, but there is no data for this above 100°C, again there is a critical need to measure this under real fuel cell conditions. Much work is being

undertaken in precious metal alloy catalysts where Pt is combined with one or more other metals and in non-precious metal catalysts [77]. These new catalytic materials are being studied in aqueous acid or in MEAs at 100% RH, very little data exists on how these materials will behave under hot and dry conditions. In fact the development of new catalyst for fuel cells run under hot and dry conditions may require their optimization outside of aqueous or water-saturated systems.

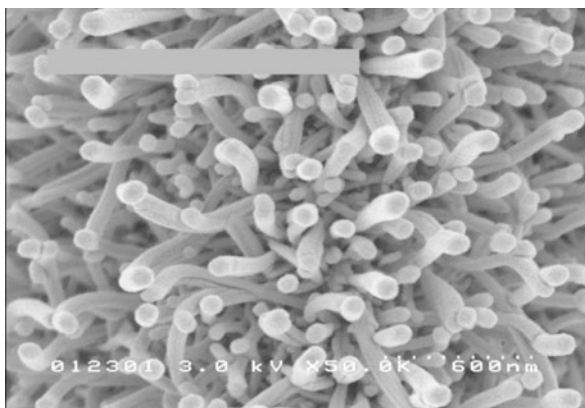
H<sub>2</sub> can be produced from fossil hydrocarbons such as natural gas or renewable biomass via reforming to produce syn gas (H<sub>2</sub> + carbon monoxide, CO), which can be converted to a H<sub>2</sub>-rich gas via the water gas shift reaction. These processes, while commonplace in chemical engineering practice, do not produce pure H<sub>2</sub>, the last 1,000 ppm or so of CO must be removed by expensive and/or inefficient unit operations such as partial oxidation, pressure swing adsorption, or membrane technology. The great advantage for the fuel cell electrodes in terms of high temperature operation is on the anode where the effect of adsorbed contaminants that slow the HOR are mitigated. This allows the anode to operate with H<sub>2</sub> contaminated with increasing levels of CO, a by-product of hydrocarbon reforming. At 80°C, CO levels as low as 10 ppm can cause significant degradations in performance but at 130°C, the fuel cell anode can tolerate 1,000 ppm of CO allowing the cost of H<sub>2</sub> produced from hydrocarbons to be dramatically reduced.

The optimum particle size for Pt in the catalyst layer is 3–5 nm [78]. Another issue is that the Pt both agglomerates and suffers from dissolution and re-precipitation. Both processes are expected to increase at higher temperature and result in higher particle sizes, lowering the rate of the ORR [79]. As described above, in a typical electrode the precious metal catalyst is supported on a carbon support that is susceptible to corrosion. It has been shown that the carbon corrodes rapidly if the electrode is held at relatively high potentials and that the reaction is first order with respect to water vapor. Carbon corrosion obviously increases with temperature and Pt loading, Fig. 12 [80]. Doping carbon with N is expected to increase the durability of the carbon.

Another solution to both the carbon support and the ionomer contact issue is to use a Pt catalyst that has no support and is embedded in the membrane such as

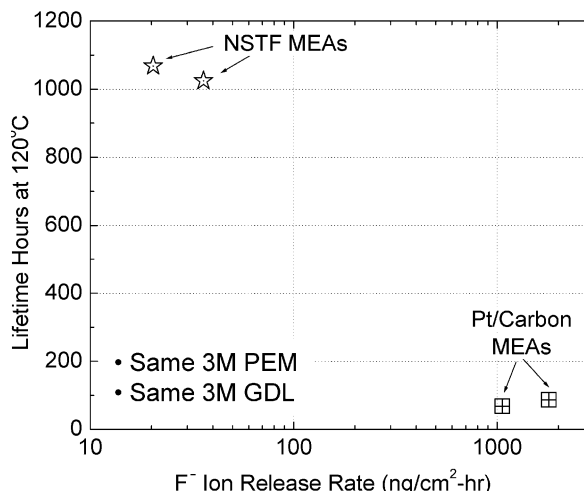


**Proton Exchange Membrane Fuel Cells: High-Temperature, Low-Humidity Operation. Figure 12** Maximum fraction of carbon consumed as a function of temperature for samples with 30–80 wt.% Pt. From reference [81]



**Proton Exchange Membrane Fuel Cells: High-Temperature, Low-Humidity Operation. Figure 13** Nano-structured thin film (NSTF) catalyst as fabricated and before transfer to a PEM. Plan view at 50,000 $\times$  original. The scale bar indicates 600 nm. From reference [5]

the Nano-structured thin film (NSTF) catalyst being developed by 3M [5, 81]. A SEM of the NSTF-Pt catalyst is shown in Fig. 13. In addition to not having a carbon support to corrode, this catalyst system is much less susceptible to Pt dissolution because the small whiskers are coated with a continuous layer of Pt, not Pt nanoparticles, and so behaves more like bulk Pt.



**Proton Exchange Membrane Fuel Cells: High-Temperature, Low-Humidity Operation. Figure 14** Hours of lifetime at 120°C (before catastrophic failure of the PEM) versus fluoride ion release rates (by IC) for NSTF and Pt/C catalyst-based membrane electrode assemblies (MEAs) having the same type PEM and GDL. Cells of 100 cm<sup>2</sup> were operated at 0.4 A/cm<sup>2</sup>, 120°C, 300 kPa, 61/84% inlet relative humidity (RH). Electrochemical surface area and crossover were measured daily at 75°C. Total lifetimes were  $\sim$ 1,800 h for the NSTF MEAs due to diagnostic testing at 75°C. End-of-life criteria were severe falloff of cell voltage and corresponding ramp-up of F<sup>-</sup> ion release indicative of membrane pinhole formation. From reference [5]

MEAs made with these electrodes also produce less F<sup>-</sup> in the effluent water coming from the fuel cell under hot and dry operating conditions, as shown in Fig. 14. The fluoride content in the effluent water coming from the cell is a common diagnostic for the rate of membrane degradation. Materials known to decompose peroxide have also been added to PEM fuel cell catalyst layer such as MnO<sub>2</sub> [82], CeO<sub>2</sub> [83], and HPA [84], and all have shown a decrease in fluoride emission rates of the fuel cell under hot and dry conditions.

Nafion<sup>TM</sup> and other PFSA ionomers work very well in conventional PEMFC electrodes because they form a thin layer that allows both gas diffusion and proton transport. For this reason, they are still the ionomer of choice in fuel cell electrodes for hot and dry operation, although under these conditions, the PFSA used usually has a much lower EW to increase proton

conductivity at the drier conditions. While the approach of adding a lower EW PFSA ionomer works well for hot and dry operating conditions, the increased swelling and hydrophilicity at lower EW leads to serve flooding if the same fuel cell is operated under high RH. As new ionomers for hot and dry operation are developed, they must be capable of extending their proton connectivity into the electrode without a large interfacial loss due to material incompatibility at the boundary of the electrode and membrane. The consequence of a change in ion-conducting material is that proton transport to and from the catalyst layer may be compromised if conventional ionomers are employed. It is true that at the current time little work has been done to develop new, stable ionomers suitable for high temperature applications that will also allow high proton conductivity and high gas permeability in the fuel cell electrodes. One may consider using stable, lower EW version of the new ionomers, functionalizing the carbon support with suitable functional groups or developing new ionomers with higher gas permeability for use as binders in the electrode. It may be necessary to completely redesign fuel cell electrodes for high temperature, low relative humidity operation using materials that are stable to oxidation, enhance the ORR, proton conductivity, and gas permeability, while maintaining suitable electrical conductivity to maintain the three-phase boundary condition during operation.

### Future Directions

This entry is by no means comprehensive. It is intended to show important examples of the approaches being taken to address the need for new materials to allow the robust operation of fuel cells under hotter and drier conditions than possible today.

At this point, there are no membranes or ionomers commercially available that will meet both the performance and durability requirements outlined above, although much progress has been made in the development of polymer membranes, which have improved conductivity and durability under these conditions. More radical approaches to the development of new fuel cell electrolytes including the development of ionomers with a variety of different protogenic groups are being explored [85, 86]. Using imidazole, ionic liquids and other replacements for water to allow

completely dry operation is also being studied [87–89]. The next few years should see significant technical advances and the introduction of improved electrolyte membranes into the marketplace.

Optimization of the electrodes for these fuel cell systems has just started. Work has been done on the optimization of electrode structure for operation under hot, dry conditions, but less has been done to study catalysis under these conditions. Part of the reason for this is that as stated above there are no commercially available polymeric materials available for the development of new electrodes studies. It is hoped that until commercially available materials for this application become available that researchers offer to share their materials. This will, however, be insufficient as the ionomers developed for catalyst layers need different properties than ionomers developed to act as fuel cell membranes. The other major issue is that catalysts for fuel cells run under conditions of water saturation have been developed using liquid phase electrochemical methods. It will be extremely important that new catalyst for fuel cells to be operated under hot, dry conditions be developed by solid-state electrochemistry. New methods must also be developed so that electrodes containing compatible ionomers can be tested.

### Acknowledgments

We gratefully acknowledge the support of numerous coworkers and collaborators. AMH would like to acknowledge support provided by the National Science Foundation MRSEC program under Grant No. DMR-0820518 at the Renewable Energy MRSEC.

### Bibliography

1. Epping Martin K, Kopasz JP, McMurphy KW (2010) Status of fuel cells and the challenges facing fuel cell technology today. In: Herring AM, Zawodzinski TA Jr, Hamrock SJ (eds) Fuel cell chemistry and operation, ACS symposium series. American Chemical Society, Washington, DC, pp 1–13
2. Ohnsman A (2010) Toyota targets \$50,000 price for first hydrogen car. Bloomberg Businessweek, 6 May 2010. <http://www.businessweek.com/news/2010-05-06/toyota-targets-50-000-price-for-first-hydrogen-car-update2.html>
3. Kranz R (2010) Industry expects fuel cell on U.S. highways in 2015. Automotive News, p 20H
4. Chen H, Cong NC, Yang W, Tan C, Li Y, Ding Y (2009) Progress in electrical energy storage systems: a critical review. Prog Nat Sci 19:291–312

5. Debe MK, Schmoeckel AK, Vernstrom GD, Atanasoski R (2006) High voltage stability of nanostructured thin film catalysts for PEM fuel cells. *J Power Sources* 161:1002–1011
6. Wainwright JS, Litt MH, Savinell RF (2003) High temperature membranes. In: Vielstien W, Gasteiger HA, Lamm A (eds) *Handbook of fuel cells: fundamentals*, vol 3, Technology and applications. Wiley, West Sussex, pp 436–446
7. Savinell R, Yeager E, Tryk D, Landau U, Wainwright J, Weng D, Lux K, Litt M, Rogers C (1994) A polymer electrolyte for operation at temperatures up to 200°C. *J Electrochem Soc* 141: L46–L48
8. Mader J, Xiao L, Schmidt TJ, Benicewicz BC (2008) Polybenzimidazole/acid complexes as high-temperature membranes. *Adv Polym Sci* 216:63–124
9. Pinery M, Eisenberg A (1987) Structure and properties in ionomers, vol 198, NATO Advanced Study Institute Series. D. Reidel, Dordrecht
10. Bazuin CG, Eisenberg A (1981) Ion-containing polymers: ionomers. *J Chem Educ* 58:938–943
11. Laconti AB, Hamdan M, McDonald RC (2003) Mechanisms of membrane degradation. In: Vielstien W, Gasteiger HA, Lamm A (eds) *Handbook of fuel cells: fundamentals*, vol 3, Technology and applications. Wiley, West Sussex, pp 647–662
12. Doyle M, Rajendran G (2003) Perfluorinated membranes. In: Vielstien W, Gasteiger HA, Lamm A (eds) *Handbook of fuel cells: fundamentals*, vol 3, Technology and applications. Wiley, West Sussex, pp 351–395
13. Hamrock SJ, Yandrasits MA (2006) Proton exchange membranes for fuel cell applications. *Polym Rev* 46:219–244
14. Maritz KA, Moore RB (2004) The state of understanding of nafion. *Chem Rev* 104:4535–4585
15. Mittelsteadt CK (2010) U.S. department of energy hydrogen program 2010 annual merit review proceedings [http://www.hydrogen.energy.gov/pdfs/review10/fc036\\_mittelsteadt\\_2010\\_o\\_web.pdf](http://www.hydrogen.energy.gov/pdfs/review10/fc036_mittelsteadt_2010_o_web.pdf)
16. Paddison SJ, Paul R (2002) The nature of proton transport in fully hydrated nafion. *Phys Chem Chem Phys* 4:1158–1163
17. Emery M, Frey M, Guerra M, Haugen G, Hintzer K, Lochhaas KH, Pham P, Pierpont D, Schaberg M, Thaler A, Yandrasits M, Hamrock S (2007) The development of new membranes for proton exchange membrane fuel cells. *ECS Trans* 11:3–14
18. Yandrasits MA, Hamrock SJ (2010) Membranes for PEM fuel cells. In: Herring AM, Zawodzinski TA Jr, Hamrock SJ (eds) *Fuel cell chemistry and operation*, ACS symposium series. American Chemical Society, Washington, DC, pp 15–29
19. Faure S, Cornet N, Gebel G, Mercier R, Pineri M, Sillion B, (1997) Sulfonated polyimides as novel proton exchange membranes for H<sub>2</sub>/O<sub>2</sub> fuel cells. In: *Proceedings of the second international symposium on new materials for fuel cell and modern battery systems*, Montreal, pp 818–825
20. Kreuer KD (1997, 2001) On the development of proton conducting materials/polymer membranes for technological applications. *Solid State Ionics* 97:1–15; Hydrogen and methanol fuel cells. *J Membr Sci* 185:29–39
21. Noshay LM, Robeson J (1976) Sulfonated polysulfone. *J Appl Polym Sci* 20:1855–1903
22. Guo Q, Pintauro PN, Tang H, O'Conner S (1999) Sulfonated and cross-linked polyphosphazine-based proton-exchange membranes. *J Membr Sci* 154:175–181
23. Buchi FN, Gupta B, Halim J, Haas O, Scherer GG (1994) A new class of partially fluorinated fuel cell membranes. *Proc Electrochem Soc* 23:220–235
24. Hickner MA, Ghassemi H, Kim YS, Einsla BR, McGrath JE (2004) Alternative polymer systems for proton exchange membranes (PEM's). *Chem Rev* 104:4587–4612
25. Sethuraman VA, Weidner JW, Haug AT, Protsailo LV (2008) Durability of perfluorosulfonic acid and hydrocarbon membranes: effect of humidity and temperature. *J Electrochem Soc* 155:B119–B124
26. Kim YS, Pivovar BS (2009) Comparing proton conductivity of polymer electrolytes by percent conducting volume. *ECS Trans* 25:1425–1431
27. King JF (1991) Acidity. In: Patai S, Rappoport Z (eds) *The chemistry of sulphonic acids, esters and their derivatives*. Wiley, New York, p 249
28. Iley J (1991) Rearrangements. In: Patai S, Rappoport Z (eds) *The chemistry of sulphonic acids, esters and their derivatives*. Wiley, New York, p 453
29. Schuster M, Kreuer KD, Andersen HT, Maier J (2007) Sulfonated poly(phenylene sulfone) polymers as hydrolytically and thermooxidatively stable proton conducting ionomers. *Macromolecules* 40:598–607
30. de Araujo CC, Kreuer KD, Schuster M, Portale G, Mendil-Jakani H, Gebel G, Maier J (2009) Poly(p-phenylene sulfone)s with high ion exchange capacity: ionomers with unique microstructural and transport features. *Phys Chem Chem Phys* 11:3305–3312
31. Litt M, Granados-Focil S, Kang J (2008) Rigid rod polyelectrolytes with frozen-in free volume: high conductivity at low RH. In: Herring AM, Zawodzinski TA Jr, Hamrock SJ (eds) *Fuel cell chemistry and operation*, ACS symposium series. American Chemical Society, Washington, DC, pp 49–63
32. Higashihara T, Matsumoto K, Ueda M (2009) Sulfonated aromatic hydrocarbon polymers as proton exchange membranes for fuel cells. *Polymer* 50:5341–5357
33. Ghassemi H, McGrath JE, Zawodzinski TA (2006) Multiblock sulfonated-fluorinated poly(arylene ether)s for a proton exchange membrane fuel cell. *Polymer* 47:4132–4139
34. Roy A, Hickner MA, Yu X, Li Y, Glass TE, McGrath JE (2006) Influence of chemical composition and sequence length on the transport properties of proton exchange membranes. *J Polym Sci B Polym Phys* 44:2226–2239
35. Penner RM, Martin CR (1985) Ion transporting composite membranes. *J Electrochem Soc* 132:514–515
36. Cleghorn S, Kolde J, Liu W (2003) Catalyst coated composite membranes. In: Vielstien W, Gasteiger HA, Lamm A (eds) *Handbook of fuel cells: fundamentals*, vol 3, Technology and applications. Wiley, West Sussex, pp 566–575

37. Tang Y, Kusoglu A, Karlsson AM, Santare MH, Cleghorn S, Johnson WB (2008) Mechanical properties of a reinforced composite polymer electrolyte membrane and its simulated performance in PEM fuel cells. *J Power Sources* 175:817–825
38. Choi J, Lee KM, Wycisk R, Pintauro PN, Mather PT (2008) Nanofiber network ion-exchange membranes. *Macromolecules* 41:4569–4572
39. Pintauro P (2009) U.S. department of energy hydrogen program 2009 annual merit review proceedings. [http://www.hydrogen.energy.gov/pdfs/review09/fc\\_09\\_pintauro.pdf](http://www.hydrogen.energy.gov/pdfs/review09/fc_09_pintauro.pdf)
40. Kerres JA (2005) Blended and cross-linked ionomer membranes for application in membrane fuel cells. *Fuel Cells* 5:230–247
41. Yang Y, Holdcroft S (2005) Synthetic strategies for controlling the morphology of proton conducting polymer membranes. *Fuel Cells* 5:171–186
42. Mao SS, Hamrock SJ, Ylitalo DA (2000) US Patent 6,090,895 crosslinked ion conductive membranes
43. Koppel IA, Taft RW, Anvia F, Zhu SZ, Hu LQ, Sung KS, DesMarteau DD, Yagupolskii LM, Yagupolski YL, Ignat'ev V, Kondratenko NV, Volkonskii AY, Slasov VM, Notario R, Maria PC (1994) The gas-phase acidities of very strong neutral Bronsted acids. *J Am Chem Soc* 116:3047–3057
44. Gubler L, Gürsel SA, Scherer GG (2005) Radiation grafted membranes for polymer electrolyte fuel cells. *Fuel Cells* 5:317–335
45. Sauguet L, Ameduri B, Boutevin B (2006) Fluorinated, crosslinkable terpolymers based on vinylidene fluoride and bearing sulfonic acid side groups for fuel-cell membranes. *J Polym Sci A Polym Chem* 44:4566–4578
46. Yandrasits MA, Hamrock SJ, Grootaert WM, Guerra MA, Jing N (2006) US Patent 7,074,841 polymer electrolyte membranes crosslinked by nitrile trimerization
47. Yandrasits MA, Hamrock SJ, Hintzer K, Thaler A, Fukushi T, Jing N, Lochhaas KH (2007) US Patent 7,265,162 bromine, chlorine or iodine functional polymer electrolytes crosslinked by e-beam
48. Desmarteau DD (1995) Novel perfluorinated ionomers and ionenes. *J Fluorine Chem* 72:203–208
49. Hamrock SJ (2010) U.S. department of energy hydrogen program 2010 annual merit review proceedings. [http://www.hydrogen.energy.gov/pdfs/review10/fc034hamrock\\_2010\\_o\\_web.pdf](http://www.hydrogen.energy.gov/pdfs/review10/fc034hamrock_2010_o_web.pdf)
50. Schaberg MS, Abulu J, Haugen GM, Emery M, O'Conner SJ, Xiong PN, Hamrock SJ (2010) New multi acid side-chain ionomers for proton exchange membrane fuel cells. *ECS Trans* 33:627–633
51. Herring AM (2006) Inorganic polymer composite membranes for proton exchange membrane fuel cells. *Polym Rev* 46: 245–296
52. Iler R (1979) *The chemistry of silica*. Wiley, New York
53. Mauritz KA, Hassan MK (2007) Nanophase separated perfluorinated ionomers as sol-gel polymerization templates for functional inorganic oxide nanoparticles. *Polym Rev* 47:543–565
54. Abbaraju RR, Dasgupta N, Virkar AV (2008) Composite nafion membranes containing nanosize TiO<sub>2</sub>/SnO<sub>2</sub> for proton exchange membrane fuel cells. *J Electrochem Soc* 155: B1307–B1313
55. Kreuer K-D, Paddison SJ, Spohr E, Schuster M (2004) Transport in proton conductors for fuel-cell applications: simulations, elementary reactions, and phenomenology. *Chem Rev* 104:4637–4678
56. Alberti G, Casciola M (2003) Composite membranes for medium-temperature PEM fuel cells. *Annu Rev Mater Res* 33:129–154
57. Malhotra S, Datta R (1997) Membrane-supported nonvolatile acidic electrolytes allow higher temperature operation of proton-exchange membrane fuel cells. *J Electrochem Soc* 144: L23–L26
58. Meng F, Aieta NV, Dec SF, Horan JL, Williamson D, Frey MH, Pham P, Turner JA, Yandrasits MA, Hamrock SJ, Herring AM (2007) Structural and transport effects of doping perfluoro-sulfonic acid polymers with the heteropoly acids, H<sub>3</sub>PW<sub>12</sub>O<sub>40</sub> or H<sub>4</sub>SiW<sub>12</sub>O<sub>40</sub>. *Electrochim Acta* 53:1372–1378
59. Alberti G, Casciola M, Capitani D, Donnadio A, Narducci R, Pica M, Sganappa M (2007) Novel Nafion-zirconium phosphate nanocomposite membranes with enhanced stability of proton conductivity at medium temperature and high relative humidity. *Electrochim Acta* 52:8125–8132
60. Coms FD, Han Liu H, Owejan JE (2008) Mitigation of perfluoro-sulfonic acid membrane chemical degradation using cerium and manganese ions. *ECS Trans* 16:1735–1747
61. Trogadas P, Parrondo J, Ramani V (2008) Degradation mitigation in polymer electrolyte membranes using cerium oxide as a regenerative free-radical scavenger. *Electrochem Solid State Lett* 11:B113–B116
62. Frey MH, Hamrock SJ, Haugen GM, Pham PT (2009) US Patent 7,572,534 fuel cell membrane electrode assembly
63. Watanabe M, Uchida H, Seki Y, Emori M, Stonehart P (1996) Self-humidifying polymer electrolyte membranes for fuel cells. *J Electrochem Soc* 143:3847–3852
64. Endoh E, Hommura S, Terazono S, Widjaja H, Anzai J (2007) Degradation mechanism of the PFSA membrane and influence of deposited Pt in the membrane. *ECS Trans* 11:1083–1091
65. Cipollini NE (2007) Chemical aspects of membrane degradation. *ECS Trans* 11:1071–1082
66. Haugen GM, Meng F, Aieta NV, Horan JL, Kuo M-C, Frey MH, Hamrock SJ, Herring AM (2007) The effect of heteropoly acids on stability of PFSA PEMs under fuel cell operation. *Electrochem Solid State Lett* 10:B51–B55
67. Liu H, Coms FD, Zhang J, Gasteiger HA, LaConti AB (2009) Chemical degradation: correlations between electrolyzer and fuel cell findings. In: Büchi FN, Inaba M, Schmidt TJ (eds) *Chemical degradation: correlations between electrolyzer and fuel cell findings polymer electrolyte fuel cell durability*. Springer, New York, pp 71–117
68. Friedmann R, Van Nguyen T (2010) Optimization of the micro-structure of the cathode catalyst layer of a PEMFC for two-phase flow. *J Electrochem Soc* 157:B260–B265



69. Litster S, McLean G (2004) PEM fuel cell electrodes. *J Power Sources* 130:61–76
70. Shao Y, Yin G, Wang Z, Gao Y (2007) Proton exchange membrane fuel cell from low temperature to high temperature: material challenges. *J Power Sources* 167:235
71. Zhang J, Xie Z, Zhang J, Tang Y, Song C, Navessin T, Shi Z, Song D, Wang H, Wilkinson DP, Liu ZS, Holdcroft SJ (2006) High temperature PEM fuel cells. *Power Sources* 160:872
72. Xu H, Song Y, Kunz HR, Fenton JM (2005) Effect of elevated temperature and reduced relative humidity on ORR kinetics for PEM fuel cells. *J Electrochem Soc* 152: A1828–A1836
73. Parthasarathy A, Srinivasan S, Appleby AJ, Martin CR (1992) Temperature dependence of the electrode kinetics of oxygen reduction at the Platinum/Nafion<sup>®</sup> Interface: a microelectrode investigation. *J Electrochem Soc* 139:2530
74. Chen YX, Li MF, Liao LW, Xu J, Ye S (2009) A thermostatic cell with gas diffusion electrode for oxygen reduction reaction under fuel cell relevant conditions. *Electrochem Commun* 11:1434–1436
75. Kucernak AR, Toyoda E (2008) Studying the oxygen reduction and hydrogen oxidation reactions under realistic fuel cell conditions. *Electrochem Commun* 10:1728–1731
76. Gasteiger HA, Kocha SS, Sompalli B, Wagner FT (2005) Activity benchmarks and requirements for Pt, Pt-alloy, and non-Pt oxygen reduction catalysts for PEMFCs. *Appl Catal B* 56:9–35
77. Thompsett D (2003) Pt alloys as oxygen reduction catalysts. In: Vielstien W, Gasteiger HA, Lamm A (eds) *Handbook of fuel cells: fundamentals, vol 3, Technology and applications*. Wiley, West Sussex, pp 467–480
78. Tada T (2003) High dispersion catalysts including novel carbon supports. In: Vielstien W, Gasteiger HA, Lamm A (eds) *Handbook of fuel cells: fundamentals, vol 3, Technology and applications*. Wiley, West Sussex, pp 481–488
79. Shao Y, Yin G, Gao Y (2007) Understanding and approaches for the durability issues of Pt-based catalysts for PEM fuel cell. *J Power Sources* 171:558
80. Stevens DA, Dahn JR (2005) Thermal degradation of the support in carbon-supported platinum electrocatalysts for PEM fuel cells. *Carbon* 43:179–188
81. Debe MK, Schmoekkel AK, Hendricks SM, Vernstrom GD, Haugen GM, Atanasoski RT (2006) Durability aspects of nanostructured thin film catalysts for pem fuel cells. *ECS Trans* 1:51–66
82. Trogadas P, Ramani V (2007) Pt/C/MnO<sub>2</sub> hybrid electrocatalysts for degradation mitigation in polymer electrolyte fuel cells. *J Power Sources* 174(1):159–163
83. Trogadas P, Parrondo J, Ramani V (2008) Degradation mitigation in polymer electrolyte membranes using cerium oxide as a regenerative free-radical scavenger. *Electrochem Solid State Lett* 11(7):B113–B116
84. Brooker RP, Baker P, Kunz HR, Bonville LJ, Parnas R (2009) Effects of silicotungstic acid addition to the electrodes of polymer electrolyte membrane fuel cells. *J Electrochem Soc* 156:B1317–B1321
85. Paddison SJ, Kreuer KD, Maier J (2006) About the choice of the protogenic group in polymer electrolyte membranes: Ab initio modelling of sulfonic acid, phosphonic acid, and imidazole functionalized alkanes. *Phys Chem Chem Phys* 8:4530–4542
86. Horan JL, Genupur A, Ren H, Sikora BJ, Kuo MC, Meng F, Dec SF, Haugen GM, Yandrasits MA, MA HSJ, Frey MH, Herring AH (2009) Copolymerization of divinylsilyl-11-silicotungstic acid with butyl acrylate and hexanediol diacrylate: synthesis of a highly proton-conductive membrane for fuel-cell applications. *ChemSusChem* 2:226–229
87. Doyle M, Choi SK, Proulx G (2000) High-temperature proton conducting membranes based on perfluorinated ionomer membrane-ionic liquid composites. *J Electrochem Soc* 147:34
88. Zhou Z, Li S, Zhang Y, Liu M, Li W (2005) Promotion of proton conduction in polymer electrolyte membranes by 1*H*-1,2,3-triazole. *J Am Chem Soc* 127:10824–10825
89. Jia L, Nguyen D, Halley JW, Pham P, Lamanna W, Hamrock S (2009) Proton transport in HTFSI–TFSI–EMI mixtures: experiment and theory. *J Electrochem Soc* 156:B136–B151

## Pulverized Coal-Fired Boilers and Pollution Control

DAVID K. MOYEDA

GE Energy, Irvine, CA, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Air Pollutant Emissions from Coal Combustion

Particulate Matter Control Technologies

SO<sub>2</sub> Emissions Control Technologies

NO<sub>x</sub> Emissions Control Technologies

Future Directions

Bibliography

### Glossary

**Anthracite** Coal which typically contains 86–97% carbon. Anthracite is considered the highest rank of coal as it has the highest energy content of all coals.

**Ash** Inorganic residues remaining after combustion.

**Baghouse** See fabric filter.

**Bituminous coal** Coal which typically contains 45–86% carbon. Bituminous coal lies between subbituminous coal and anthracite in terms of rank, and

is commonly divided into additional subgroups dependent upon the content of volatile material.

**Calorific value** Corresponds to the amount of heat per unit mass when combusted. Can be expressed as gross calorific value, which is the amount of heat liberated during combustion under standardized conditions at constant volume so that all of the water in the products remains in liquid form, or as net calorific value, which is the maximum achievable heat release obtainable in a furnace at constant pressure.

**Carbon dioxide (CO<sub>2</sub>)** A heavy, colorless gas that results from the combustion of fossil fuels and from natural sources.

**Carbon monoxide (CO)** A colorless, odorless gas produced by incomplete combustion of fossil fuels.

**Coal** A solid fossil fuel consisting primarily of carbon, hydrogen, nitrogen, oxygen, sulfur, and nitrogen. Coal also contains ash, minerals which do not burn, and moisture. Coal is typically classified or ranked by its volatile matter, fixed carbon content, and calorific value.

**Dry FGD** A process that removes sulfur oxides from the flue gas and results in the formation of a dry product or waste.

**Electrostatic precipitator (ESP)** A device for removing particulate from a gas stream based upon using an electric field to charge the particles in the gas and move them to a collecting surface.

**Fabric filter** A device for removing particulate from a gas stream based upon filtering the gas through a filter media.

**Flue gas desulfurization (FGD)** Technologies that are used to remove sulfur oxides from the flue gas.

**Lignite** Coal which contains 25–35% carbon and which has a lower calorific value than subbituminous and bituminous coals and typically higher moisture and volatile content. Lignite is the lowest range of coal.

**Low-NO<sub>x</sub> burners (LNB)** Technology for reducing NO<sub>x</sub> emissions by controlling fuel and air mixing in the flame.

**Nitric oxide (NO)** A colorless gas resulting from the combustion of fossil fuels.

**Nitrogen dioxide (NO<sub>2</sub>)** A reddish-brown gas that can be emitted from the combustion of fossil fuels or is formed by atmospheric reaction of nitric oxide (NO) and oxygen (O<sub>2</sub>).

**NO<sub>x</sub>** Refers to the total nitric oxide (NO) and nitrogen dioxide NO<sub>2</sub> concentration.

**Overfire air (OFA)** Technology that reduces NO<sub>x</sub> emissions based upon air staging.

**Reburning** Technology that reduces NO<sub>x</sub> emissions based upon staging fuel in a fashion that permits fuel fragments to reduce (or reburn) nitric oxide (NO) in the flue gas.

**Selective catalytic reduction (SCR)** Technology that reduces NO<sub>x</sub> emissions by mixing ammonia into the flue gas and reacting the ammonia with NO<sub>x</sub> over a catalyst.

**Selective noncatalytic reduction (SNCR)** Technology that reduces NO<sub>x</sub> emissions by mixing an amine-based reagent into the flue gas at a temperature which selectively promotes the reaction of amine (NH<sub>2</sub>) with nitric oxide to form molecular nitrogen (N<sub>2</sub>)

**Subbituminous coal** Coal which typically contains 35–45% carbon and which typically has a lower calorific value than bituminous coal and higher moisture and volatile content.

**Sulfur dioxide (SO<sub>2</sub>)** A colorless, irritating gas resulting from the combustion of sulfur contained in fossil fuels, particularly coal.

**Sulfur oxides** Refers to sulfur dioxide (SO<sub>2</sub>) and sulfur trioxide (SO<sub>3</sub>).

**Volatile matter** Non-moisture component of coal that is liberated at high temperature in the absence of air.

**Wet FGD** A process that removes sulfur oxides from the flue gas and results in the formation of a product or waste that is a solution or slurry.

## Definition of the Subject

Fossil fuels, such as coal, natural gas, and fuel oil, are used to generate electric power for industrial, commercial, and residential use. Due to its relatively low cost and abundance throughout the world, coal has historically played a significant role in energy production and approximately 41% of the world power generation was supplied by coal-fired power plants in 2008 [1]. While increased discoveries of natural gas and fuel oil resources, and growth in renewable energy sources, such as wind, solar, and geothermal energy is projected to reduce the use of coal for power generation, energy

from coal will continue to be used to satisfy the world's energy demands.

One drawback in the use of coal for power production is that it produces high levels of air pollutants, such as particulate, sulfur oxides, and nitrogen oxides. Coal also produces higher carbon dioxide emissions than other fossils such as natural gas and fuel oil. In addition, due to the large quantities of coal that are used for power production, emissions of toxic metals, such as mercury, which are contained in very small quantities in the coal, can also be a concern.

Over the past several decades, a number of technologies have been developed to reduce the air pollutant emissions formed from coal combustion. Modern power plants can be equipped with advanced technologies that reduce particulate, sulfur oxide, and nitrogen oxide emissions to the most stringent levels. These technologies are the focus of this entry.

## Introduction

The abundance of coal throughout the world led to its use in China as early as 1000 B.C. and by the Romans in Britain before 400 A.D. [2]. While the use of coal in Britain largely disappeared when the Romans left in the fifth century, coal use in England increased in the thirteenth century, and by the beginning of the seventeenth century, coal was the dominant source of energy [3]. During the industrial revolution of the eighteenth and nineteenth century, the invention and development of the steam engine led to an increase in coal production and use for industrial processes and transportation [4–6]. In the late nineteenth century, advances in electricity and the invention of a reliable incandescent lamp set the stage for the use of coal to generate electrical power with the first coal-fired central generating plant in the USA established in 1882 [7, 8]. The growing demand for electrical power led to further developments in steam-generating boiler technology, such as pulverizing the coal prior to introducing it to the boiler furnace. Pulverized coal firing enabled the construction of larger boilers and power plants and became the predominant firing method for large steam-generating power plants beginning in the late 1920s [9, 10].

The air pollution associated with coal combustion was recognized in England as early as the thirteenth

century, where the burning of coal in urban areas created a smoky environment and led to bans on coal use [11]. While improvements in combustion methods and the use of chimneys or stacks to disperse the smoke overcame some of the objections to coal burning, fundamentally, the growing need for low-cost energy offset public concerns over the unhealthy aspects of coal combustion, and such bans were largely ignored [3]. As the world entered the twentieth century, widespread industrialization and the increased use of coal resulted in degraded air quality in London and other English cities [12]. By the middle of the century, severely degraded air quality in Los Angeles, California [13], and air quality disasters in London, England [14, 15], and Donora, Pennsylvania [16], heightened public awareness of the dangers of air pollution and government recognition of the need for research to understand the formation of and problems associated with air pollutants and for the development of regulations to limit their emissions [17].

For coal combustion, the primary air pollutants of concern are particulate matter (PM), sulfur dioxide ( $\text{SO}_2$ ), and oxides of nitrogen ( $\text{NO}$  and  $\text{NO}_2$ , which are referred to as  $\text{NO}_x$ ) [18]. These pollutants originate from the ash, sulfur, and nitrogen species present in the coal. Particulate matter reduces visibility and contributes to regional haze. In addition, “coarse” particles (from 2.5 to 10  $\mu\text{m}$  in diameter,  $\text{PM}_{10}$ ) and “fine” particles (smaller than 2.5  $\mu\text{m}$  in diameter,  $\text{PM}_{2.5}$ ) can accumulate within different areas of the respiratory system and aggravate health problems such as asthma or lead to increased respiratory symptoms and disease [19–21]. Exposure to sulfur and nitrogen oxides in sufficient concentration can lead to respiratory symptoms, particularly for those susceptible to these problems [22, 23]. In addition, sulfur and nitrogen oxide emissions can react with water in the atmosphere to form acids that deposit in lakes and soils leading to acidification [24–26], which is referred to as “acid rain.” Nitrogen oxides also react with volatile organic compounds to form photochemical oxidants, such as ozone ( $\text{O}_3$ ), which present a hazard to human health and plants in high concentrations and which cause visible “smog” in urban areas [27, 28].

This article will introduce how PM,  $\text{SO}_2$ , and  $\text{NO}_x$  emissions are formed during coal combustion and will discuss the technologies that have been developed to

control these emissions from pulverized coal-fired power plants. The technology discussion will focus on the primary technologies that have been applied and are available on a commercial scale.

### Air Pollutant Emissions from Coal Combustion

The primary air pollutants regulated from coal-fired power plants worldwide are carbon monoxide (CO), sulfur dioxide (SO<sub>2</sub>), oxides of nitrogen (NO and NO<sub>2</sub>, which are referred to as NO<sub>x</sub>), and particulate matter (PM) [29]. In general, CO emissions from pulverized coal-fired power boilers are low (<50–200 ppmv) as the combustion system tends to be operated with sufficient excess air to maximize combustion efficiency [30]. Local standards, rather than national standards, are generally established to limit CO emissions and can vary widely. SO<sub>2</sub> and NO<sub>x</sub> emissions are regulated as both can be a health hazard in high concentrations [22, 23] and are contributors to dry and wet acid deposition [24–26]. In addition, NO<sub>x</sub> emissions contribute to ground-level ozone formation through the reaction with volatile organic compounds [27, 28]. Particulate emissions can lead to reduced visibility and degraded air quality. Small particulates emitted from combustion (<10 μm) or formed by reactions of SO<sub>2</sub> and NO<sub>x</sub> in the environment (<2.5 μm) are associated with increased respiratory symptoms and disease [19–21].

In addition to these primary pollutants, since the early 1990s, there has been significant interest in the emissions of hazardous air pollutants (HAPs), such as toxic organic compounds, acid gases, and metals, from coal combustion. Based upon comprehensive studies performed to characterize HAPs from fossil fuel-fired power plants [31, 32], the US Environmental Protection Agency (EPA) identified mercury as the primary HAP of concern due to the contribution of coal-fired power plants to the total anthropogenic emissions of this compound and to the risk of exposure to methylmercury through the consumption of contaminated fish [33]. Methylmercury is a highly toxic neurotoxin that bioaccumulates up the food chain [34]. The concern over mercury emissions from coal combustion is shared in other developed areas of the world [35–37]. In addition to mercury, other HAPs of concern from coal-fired power plants due to their levels of emissions

and potential health risks include carcinogenic metals, such as chromium, nickel, and arsenic, and acid gases, including hydrogen chloride (HCl) and hydrogen fluoride (HF), which are soft-tissue irritants and can cause respiratory disease [38].

A by-product of the combustion of fossil fuels is the generation of carbon dioxide (CO<sub>2</sub>). In the atmosphere, CO<sub>2</sub> and other gases, such as water vapor, absorb and reemit infrared radiation reflected from the earth's surface, resulting in a "greenhouse" effect that raises surface temperature of the earth [39]. Since coal has a higher carbon-to-hydrogen ratio than other fossil fuels, such as natural gas and fuel oil, CO<sub>2</sub> emissions from coal combustion are higher than those from the combustion of other fossil fuels [40, 41]. The high carbon footprint of coal-fired power plants has led to an increased demand for the generation of power from alternative sources, such as cleaner fuels (e.g., natural gas) and renewable sources (e.g., biomass, wind, and solar), to the development of high-efficiency coal-fired power plants (e.g., integrated gasification combined cycle (IGCC)), and to the research and development of methods for carbon capture and sequestration from large coal-fired power plants [42, 43].

This article focuses on the primary pollutants from coal. The levels of SO<sub>2</sub>, NO<sub>x</sub>, and PM from coal-fired power plants vary widely and are determined by the coal type and combustion system design. Table 1 compares the major properties of several coals, low, medium, and high-sulfur bituminous coals; subbituminous coal; and lignitic coal. As can be seen in the table, the sulfur, nitrogen, and ash content contained in coals can vary significantly. Bituminous coals are characterized by a higher heating value in terms of energy release per mass of fuel consumed than lower rank subbituminous coal and lignite, which have higher moisture contents in comparison to bituminous coal. Bituminous coals also tend to have higher sulfur and nitrogen content in comparison to the low-rank coals. In the remainder of this entry, the impacts of coal ash, sulfur, and nitrogen content and boiler design and operation on emissions from coal-fired power plants will be discussed.

### Particulate Emissions

Particulate emissions from pulverized coal combustion result primarily from the mineral matter included in

Pulverized Coal-Fired Boilers and Pollution Control. Table 1 Typical coal properties

Property	Units	Bituminous low sulfur	Bituminous medium sulfur	Bituminous high sulfur	Subbituminous	Lignite
<i>Ultimate analysis</i>						
Carbon	% dry	69.36	70.06	62.08	65.54	66.15
Hydrogen	% dry	5.32	5.00	4.44	4.15	4.20
Nitrogen	% dry	1.50	1.66	1.07	0.95	0.96
Sulfur	% dry	1.04	3.08	7.40	0.79	0.37
Oxygen	% dry	12.73	7.54	6.15	14.00	20.72
Ash	% dry	10.05	12.66	18.86	14.57	7.60
<i>Proximate analysis</i>						
Volatile matter	% a.r.	42.97	38.15	34.87	31.16	27.02
Fixed carbon	% a.r.	43.21	43.10	42.79	34.88	33.38
Ash	% a.r.	9.62	11.77	18.05	11.26	4.97
Moisture	% a.r.	4.20	6.98	4.29	22.70	34.63
<i>Dry, ash-free basis</i>						
Nitrogen	% d.a.f.	1.67	1.90	1.32	1.11	1.04
Sulfur		1.16	3.53	9.12	0.92	0.40
Volatile matter		49.86	46.95	44.90	47.18	44.74
Fixed carbon		50.14	53.05	55.10	52.82	55.26
<i>Gross heating value</i>						
	kJ/kg	27,244	27,386	25,663	20,002	16,866
	Btu/lb	11,713	11,774	11,033	8,599	7,251

a.r. as received

d.a.f. dry, ash-free

the coal. The mineral matter can be classified as being inherent (chemically bound into the coal matrix), included (present in the coal matrix), or extraneous (soil and rock mixed into the coal during mining) [44]. The composition of the mineral matter varies according to the geographic location, type of coal, and how much extraneous material is entrained into the coal during mining. The primary minerals in coal are quartz, aluminosilicates, iron sulfides, and carbonates [45]. These minerals can be broken down into the major ash constituents shown in Table 2 [46]. The primary constituents are silica and alumina oxides ( $\text{SiO}_2$  and  $\text{Al}_2\text{O}_3$ ). The iron (Fe) and alkali metal (calcium (Ca), magnesium (Mg), and sodium (Na)) content vary widely but are important as these metals

have a major influence on the slagging and fouling characteristics of the resulting ash [47]. Slagging refers to the buildup of molten or partially fused ash on the furnace walls or radiant heat transfer surfaces of the boiler. Fouling refers to the deposit of ash on the convective heat transfer surfaces such as the superheater and reheater.

During combustion, the mineral matter in coal undergoes several transformations to become particulate matter that is entrained in the flue gas. The size and composition of the particulate matter depends upon the mineral composition, how it is included in the coal, the presence of other species, and the time-temperature history of the coal particles as they are being burned [48]. There are two primary mechanisms for ash

**Pulverized Coal-Fired Boilers and Pollution Control. Table 2** Major constituents in ash for typical US coals

State coal type		Pennsylvania bituminous	Utah bituminous	Wyoming subbituminous	North Dakota lignite
<i>Ash composition</i>					
SiO <sub>2</sub>	wt.%, S-free	45.3	60.0	43.3	22.0
Al <sub>2</sub> O <sub>3</sub>		24.2	22.7	17.2	20.4
Fe <sub>2</sub> O <sub>3</sub>		20.3	4.1	6.3	11.8
TiO <sub>2</sub>		1.2	1.2	1.4	0.5
O <sub>2</sub> O <sub>5</sub>		0.6	1.5	2.5	0.1
CaO		4.8	4.6	22.7	30.3
MgO		1.1	1.9	4.0	8.0
Na <sub>2</sub> O		1.4	1.1	1.7	5.1
K <sub>2</sub> O		1.3	1.8	0.5	1.4

formation during combustion. First, ash may remain with the burning coal particle, melt at high temperature and coalesce to form liquid droplets, and then agglomerate with other ash particles [49]. Fragmentation of the coal char leads to the formation of additional ash particles smaller than the parent particle [50]. Second, at high temperatures, the more volatile metals will vaporize [51]. These vapors can then undergo homogeneous nucleation to form very fine particulate that grow larger by the condensation of additional volatile species onto the particle surface or can form larger particles by coagulation and chain agglomerate formation. Ash particles formed by the first mechanism represent the bulk of the fly ash mass and are greater than 1  $\mu\text{m}$  in diameter, with a typical size range of 3–50  $\mu\text{m}$ , and those by the second mechanism are less than 0.5  $\mu\text{m}$  in diameter, with a peak around 0.1  $\mu\text{m}$ .

The total particulate emissions in the flue gas from a coal-fired boiler depend upon the design of the combustion system and the amount of unburned carbon resulting from incomplete combustion. For pulverized coal-fired boilers, 70–90% of the ash will typically end up in the flue gas (fly ash), and the remaining 10–30% of the ash will collect on the walls of the boiler and end up as ash removed from the bottom of the boiler (bottom ash) [9]. Cyclone-fired boilers and other boilers with wet bottom designs collect more bottom ash and generate approximately 15–30% fly ash. Boiler

load can also impact particulate emissions with lower boiler loads resulting in lower emission rates depending upon boiler design and fuel characteristics [52]. Periodic cleaning of heat transfer surfaces, such as the superheater, reheater, economizer, and air preheater, via soot blowing can lead to short-term increases in particulate loading in the flue gas.

### Sulfur Oxides Formation

Sulfur is present in coal in inorganic and organic forms. The primary inorganic form is as iron sulfide ( $\text{FeS}_2$ ), which is typically in pyrite (cubic) form. A small amount of sulfur is also present as inorganic sulfates, typically as salts of minerals such as calcium and iron ( $\text{CaSO}_4 \bullet 2\text{H}_2\text{O}$  and  $\text{FeSO}_4 \bullet 7\text{H}_2\text{O}$ ). The organic compounds containing sulfur are larger chain hydrocarbons and are believed to consist of thiols, sulfides, and thiophenes [53]. The proportions of inorganic and organic sulfur vary depending upon the coal; however, organic sulfur typically comprises between 30% and 50% of the total sulfur [54]. The remainder is primarily pyrite, as inorganic sulfates are typically less than 0.1% of the total sulfur.

During the combustion process, the inorganic and organic sulfur is released and converted to sulfur dioxide,  $\text{SO}_2$ . For pyrite, this process involves decomposition to  $\text{FeS}$  and subsequent oxidation [55]. For the organic sulfur, this process involves decomposition

and oxidation of the parent compound to release the sulfur-bearing species. The gas-phase sulfur chemistry is complex [56], but in fuel-lean flames, the released sulfur species are readily oxidized to  $\text{SO}_2$ . Under fuel-rich conditions, hydrogen sulfide ( $\text{H}_2\text{S}$ ) and carbonyl sulfide ( $\text{COS}$ ) are also present in low concentrations [57, 58], but these species are readily oxidized to  $\text{SO}_2$  in the high-oxygen environment of practical combustion systems.

For bituminous and subbituminous coals, nearly all of the sulfur in the coal (90–95%) is released as sulfur oxides, both  $\text{SO}_2$  and sulfur trioxide,  $\text{SO}_3$ , with the remainder ending up as sulfates in the ash. For pulverized coal-fired boilers, emissions of  $\text{SO}_2$  are independent of the boiler design and operating conditions [59]. For lignitic and subbituminous coals containing ash with a high alkali (calcium and sodium) content, a higher fraction of the sulfur is retained in the ash. Sulfur oxide emissions from lignites can be correlated with the sulfur, sodium oxide ( $\text{Na}_2\text{O}$ ), and silica ( $\text{SiO}_2$ ) content of the ash [60].

The fraction of sulfur that is emitted as  $\text{SO}_3$  is small and is typically 0.5–1.5% of the total sulfur in the coal [61]. While equilibrium favors the formation of  $\text{SO}_3$  at low temperatures, the reaction kinetics are too slow to permit appreciable  $\text{SO}_3$  to form prior to the flue gas exiting the stack [49]. The  $\text{SO}_3$  that does form is a concern as, at low temperatures, it reacts with moisture to form sulfuric acid, which can condense out of the flue gas and cause corrosion. The coal sulfur content and fly ash composition and boiler design and operating variables, such as excess oxygen and flue gas residence time-temperature profile, are all factors that can influence the concentration of  $\text{SO}_3$  that is emitted [62].

### Nitrogen Oxides Formation

During combustion,  $\text{NO}_x$  emissions can form through three main mechanisms. The first process is fixation of nitrogen in the air (atmospheric nitrogen) via the overall reaction:

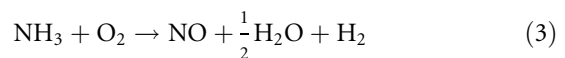
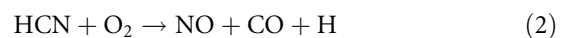


The detailed reaction is believed to be initiated by a free oxygen atom (O) attacking the very stable nitrogen molecule ( $\text{N}_2$ ) to form NO and a free nitrogen atom (N), which can then attack  $\text{O}_2$  [63]. As this

chain reaction can only be initiated at high temperatures, NO formed from this process is referred to as “thermal  $\text{NO}_x$ .” The second mechanism is through the reaction between hydrocarbons in the flame front with molecular nitrogen [64]. As this reaction can produce NO levels higher than those expected from thermal NO formation alone within the initial stages of combustion, it is referred to as “prompt  $\text{NO}_x$ .” The third mechanism is through oxidation of fixed nitrogen species present in the fuel. NO formed from this process is referred to as “fuel  $\text{NO}_x$ .” For pulverized coal combustion, the majority of the  $\text{NO}_x$  emissions result from fuel nitrogen rather than thermal  $\text{NO}_x$  formation [65, 66]. Prompt  $\text{NO}_x$  is estimated to be less than 5% [67].

The nitrogen content in coal typically varies between 1% and 2%, with some coals having nitrogen contents higher and lower than this range. Analytical techniques suggest that the majority of the nitrogen is contained in five-member (pyrrolic) or six-member (pyridinic) aromatic ring structures [68, 69], with pyrrolic nitrogen being the dominate form for all coal ranks. A smaller fraction of nitrogen is present in other species which may include quaternary nitrogen (i.e., a nitrogen molecule with four bonds) or aromatic amines [70]. The form and distribution of nitrogen within the coal matrix is expected to impact how the nitrogen is released during the combustion process and the potential for its conversion to NO.

As a coal particle devolatilizes during combustion, a fraction of the nitrogen is released with the volatiles, while a fraction of the nitrogen stays in the coal char [71]. The volatile nitrogen may be contained in tars or in the form of cyanic and amine species, which are believed to be the products of rapid secondary pyrolysis of the hydrocarbon forms. As the tars undergo further breakdown, a fraction of the nitrogen is released primarily in the form of HCN [72], and a fraction is incorporated into soot compounds [73]. The nitrogen species released during devolatilization can undergo oxidation to form NO via the overall reactions:



In the absence of oxygen, the fixed nitrogen species can be reduced to  $\text{N}_2$ . After the volatiles are released,

further nitrogen species, primarily as HCN, can be released from the coal char by thermal dissociation of the remaining organic solids. As the coal char burns, NO can be formed by heterogeneous oxidation of nitrogen remaining in the char [74].

Pilot-scale studies to evaluate the impacts of coal type and combustion conditions on the formation of NO<sub>x</sub> emissions from pulverized coal combustion have shown that as the total fuel nitrogen content increases, NO<sub>x</sub> emissions increase, but that the fraction of fuel nitrogen evolved with the volatiles also impacts NO<sub>x</sub> emissions [75]. Large differences in NO<sub>x</sub> emissions can result from coals having similar nitrogen contents and burned under the same conditions when fuel nitrogen is evolved at different rates. Fuel and air contacting also has a strong influence on NO<sub>x</sub> emissions with rapid and intimate contact of air with the fuel during devolatilization leading to higher NO<sub>x</sub> emissions than those which result from the slow mixing processes typical of a diffusion flame [76]. The staging inherent in a diffusion flame lowers NO<sub>x</sub> emissions by allowing fuel nitrogen to evolve in an oxygen-free environment where it can be reduced to N<sub>2</sub> rather than being oxidized to NO [77]. In a practical pulverized coal flame, coals which evolve fuel nitrogen early in the combustion process tend to produce lower NO<sub>x</sub> emissions than those that retain more nitrogen in the char.

The coal type and the boiler design and combustion system all impact the NO<sub>x</sub> emissions produced from pulverized coal-fired power plants [75, 78]. Low-rank coals (i.e., subbituminous and lignite) produce lower NO<sub>x</sub> emissions than high-rank coals (i.e., bituminous and anthracite). This is primarily due to the lower nitrogen and higher volatile content of the low-rank fuels. For power generation, the majority of pulverized coal fired boilers use either an array of circular burners located on one or two walls of the furnace (i.e., wall or opposed wall firing) or columns of burners arranged on the furnace corners firing in a tangential pattern (i.e., tangential firing). Tangential firing results in longer, slower mixing flames and produces lower NO<sub>x</sub> emissions than wall-fired boiler designs, which tend to have shorter, faster mixing flames. Other firing system designs, such as cyclone and arched-fired boilers, that are designed to remove the coal ash in a molten state tend to produce the highest NO<sub>x</sub> emissions due to intense fuel and air mixing and higher temperatures.

For a given boiler design, firing configuration, and coal, NO<sub>x</sub> emissions can be approximately correlated with the total heat liberation per unit of cooled surface area [79]. Older boiler designs tend to have a high ratio of heat release per surface area resulting in high NO<sub>x</sub> emissions, while more modern designs use a lower ratio to reduce NO<sub>x</sub> emissions.

### Particulate Matter Control Technologies

To remove dust from a gas stream, such as fly ash contained in boiler flue gas, a force must be applied that causes the particles to divert from the flow direction of the gas long enough for the particles to contact a collecting surface. The collecting force can be gravitational, centrifugal, inertial, direct interception, diffusional, or electrostatic [80]. Some collection devices may use a combination of these forces. The effectiveness of the collecting force depends upon the characteristics of the particulate matter to be collected; hence, the particulate matter size distribution and chemical and physical characteristics must be known to select the most appropriate collecting device [81].

Modern coal-fired power plants use two primary devices to remove and collect fly ash: electrostatic precipitators and fabric filters. Older boilers and industrial boilers may also use mechanical collectors, such as cyclones, to remove larger fly ash particles prior to the remaining particulate being collected in an electrostatic precipitator or fabric filters. Venturi scrubbers have also been used for large coal-fired boilers; however, these are less common and would not be used for compliance with current emissions regulations. The remainder of this section will provide an introduction to electrostatic precipitators and fabric filters.

### Electrostatic Precipitator

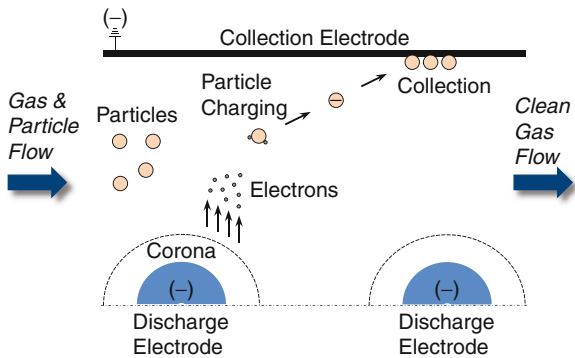
If a dust-laden gas is passed through a strong electrical field, the dust will become charged and will begin to flow in the direction of the ion flow, enabling them to be removed from the gas. To apply this phenomenon in practice requires four steps: (1) the particles need to be charged, (2) the particles need to be collected on a surface, (3) the particles need to be removed from the surface in a fashion that minimizes their reentrainment into the gas flow, and (4) the particles need to be removed from the device [82]. The collecting device



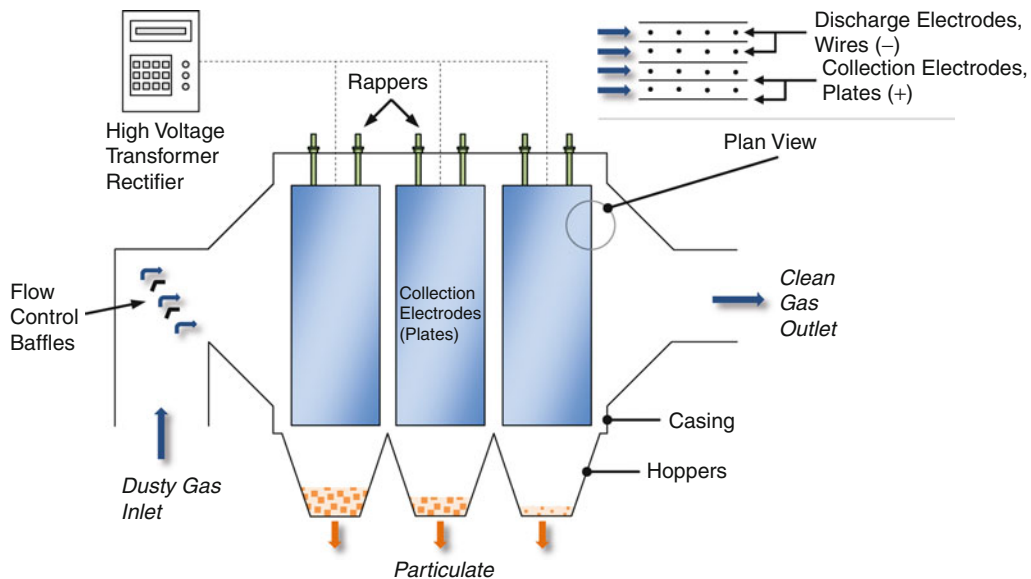
that accomplishes these steps is typically referred to as an electrostatic precipitator (ESP). As shown in Fig. 1, in an ESP, an electric potential is set up between a discharge electrode and a collection electrode, and the dust laden gas is passed between the electrodes at a relatively low velocity. When the potential is high enough, corona discharge is generated near the discharge electrode that begins to ionize the gas molecules; these molecules then migrate to the lower potential collecting electrode. In the process, the molecules

attached themselves to the dust particles which then also migrate to the electrode. Upon reaching the electrode, the particles lose their charge and stick to the electrode. As the dust layer on the collection electrode builds up, the electrical current is reduced, and it becomes necessary to vibrate the plate to remove the particles. This is accomplished through a rapper system that strikes the top of the collection plate periodically to remove the dust. The dust then falls into the collection hopper where it can be removed.

ESPs are generally classified by the method in which the particulate is removed from the collection electrode, e.g., wet versus dry, and by the geometry of the electrodes, e.g., wire in tube or wire and plates. The most common design for coal-fired power plants is the dry, wire and plate type as illustrated in Fig. 2. The main components of an ESP are a gas-tight casing or outer shell, discharge electrodes, collection electrodes, a high-voltage transformer rectifier or high-frequency power supply for application of electrical power, rappers to remove particulate from the collection electrode, and pyramidal-shaped hoppers to remove particulate from the system [83]. The collection electrodes or plates are arranged in multiple mechanical fields, where each field consists of a series of equally spaced plates perpendicular to the gas flow.



**Pulverized Coal-Fired Boilers and Pollution Control. Figure 1**  
Principle of electrostatic precipitator operation



**Pulverized Coal-Fired Boilers and Pollution Control. Figure 2**  
Overview of electrostatic precipitator

The discharge electrodes or wires are suspended in the gas passage between each pair of plates. An ESP will have one or more mechanical fields. A high-voltage pulsed direct current or direct current is applied to the discharge electrode. When the voltage is high enough, a corona discharge will be generated that causes particles to be moved to the collection electrode.

The performance of an ESP is impacted by the design of the ESP, the strength of the electric field, and the particle characteristics. The theoretical collection efficiency of an electrostatic precipitator can be expressed by the Deutsch-Anderson equation [84]:

$$\eta = 1 - e^{-(A/V)\omega} \quad (4)$$

where  $\eta$  is the collection efficiency for a given particle size,  $A$  is the surface area of the collecting electrode,  $V$  is the volumetric gas flow rate, and  $\omega$  is the migration velocity of the particle. The term  $(A/V)$  is referred to as specific collection area (SCA) of the ESP. Equation 4 shows that the collection efficiency of an ESP can be improved by either increasing the surface area of the collecting electrodes or decreasing the volume of gas to be treated. The theoretical migration velocity of the particle can be calculated by comparing the electrostatic forces on the particle to the drag force and, with simplifying assumptions, can be expressed as [85]:

$$\omega = \frac{E_c E_p d}{4\pi\mu_g} \quad (5)$$

where  $E_c$  is the charging field strength,  $E_p$  is the precipitating (collecting) field strength,  $d$  is the particle diameter, and  $\mu_g$  is the gas viscosity. Equation 5 shows that the particle migration velocity is proportional to both the electric potential and the particle size. As particle size is reduced, voltage must be increased to maintain migration velocity and collection efficiency. Equation 5 suggests that continuing to increase the electric field will continue to increase the particle migration velocity and, hence, the ESP collection efficiency. While this is true, a practical limit is reached where the field breaks down and sparking between the electrodes occurs [80]. Sparking destroys the electrical field and lowers the collection efficiency.

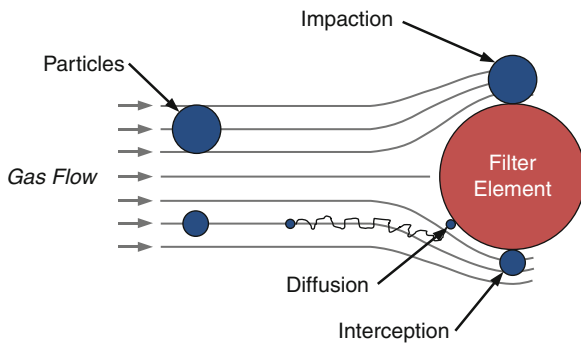
The electrical conductivity of the particulate matter has an impact on the electrical field between the discharge and collection electrodes. The resistance to

electrical conductivity is called the “resistivity” [86]. As the dust layer builds up on the collection electrode, the voltage at which sparking occurs is lowered due to the increased electrical field at the dust layer. If the resistivity of the ash is too high, charge will build up on the collected particles and can become high enough to cause an electrical breakdown which causes ions of the opposite polarity to be injected back into the gap which reduces the charge on the particles in the gas flow and which can cause sparking. This breakdown is referred to as “back corona” [87]. The resistivity of fly ash is a function of the particle composition, gas temperature, and concentrations of water vapor and  $\text{SO}_3$  in the flue gas [88].

Properly designed and operating ESPs can remove between 99% and 99.9% of the total particulate matter. Older equipment is generally less efficient. As noted above, the primary factors that influence the ESP performance are the particle resistivity, the particle size distribution, and the flue gas temperature and flow rate [83]. As these parameters vary from the conditions assumed in designing the ESP, the ESP collection efficiency can be impacted. Other factors that can impact the ESP performance are the electrical conditions in the ESP, the gas flow distribution, and the particle reentrainment during rapping [89]. The applied voltage in each field of the ESP needs to be optimized to maximize the acceptable spark rate with the specific ash characteristics. The gas flow distribution entering the ESP should be optimized to ensure a uniform flow distribution between the collecting plates, and the ESP should be designed to minimize the potential for gas flow to bypass the plates (sneakage). The rapping system design and frequency should be optimized to minimize the reentrainment of particles into the gas stream.

### Fabric Filter Baghouse

If a dust-laden gas is passed through a fiber bed or filter, the dust will be removed from the gas and will collect on the filter. The principle mechanisms of filtration include impaction, interception, and diffusion, which are related to the relative particle and fiber size and velocity through the filter [90]. These mechanisms are illustrated in Fig. 3. Particles will impact on a fiber when they are too large to follow the gas streamlines



### Pulverized Coal-Fired Boilers and Pollution Control.

**Figure 3**

Principal mechanisms involved in fabric filtration

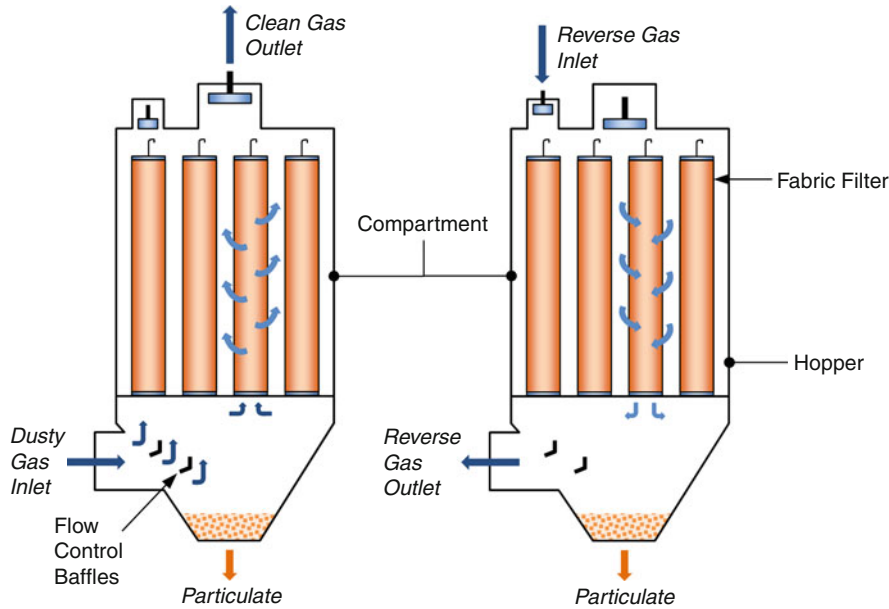
around the filter. Interception will occur when a particle follows a streamline close to a fiber and is attracted to the surface by van der Waals forces. Particles that are very small are influenced by Brownian diffusion and can come into contact with the fiber and be removed from the gas. Other forces, such as electrostatic and gravity, can also play a role in removing particulate from the gas.

The filtration media can be flat and supported in a frame or in bags which are supported on cages. As bags are most common for large systems, the containing device is commonly called a baghouse. In a baghouse or fabric filter, the dust-laden gas passes through the suspended filtration media. Particles impact on the filter and are held. As collection proceeds, a deposit begins to build up that also serves as a means of collecting particulate. Eventually, the deposit must be removed or the pressure drop will be too high. As this can be accomplished by several means, baghouses are most commonly classified by the cleaning method [91]. Typical cleaning methods include mechanical, reverse airflow, and pulse-jet cleaning. Mechanical cleaning typically involves flowing the gas on the inside of the bag, stopping the gas flow, and shaking the bag to remove particulate. Reverse airflow cleaning consists of periodically flowing gas in the opposite direction to the normal gas flow to remove the particulate buildup from the filter. In a pulse-jet system, the gas flows from the outside to the inside of the bag. Periodically, a pulse of compressed air is injected down the center of the bag to flex the bag and remove particulate.

Reverse airflow and pulse-jet baghouses have been applied to coal-fired boilers throughout the world [92, 93]. Simplified overviews of the cleaning methodology and main components for reverse airflow and pulse-jet fabric filters are illustrated in Figs. 4 and 5. The similar components for each type of fabric filter include a gas-tight casing, fabric filters, supporting frames, and pyramidal-shaped hoppers to remove particulate from the unit. Large-scale systems will be designed with multiple compartments to facilitate cleaning and maintenance. Reverse airflow baghouses include a fan and ductwork and valves for flowing cleaned gas back through the unit. Pulse-jet baghouses include a compressed air header and distribution plenum to periodically introduce compressed air into the throat of the filter bags. While significant experience exists with the use of reverse airflow baghouses on large utility boilers, interest in pulse-jet baghouses is increasing since this design uses higher air-to-cloth ratios than the reverse airflow design, resulting in smaller equipment size and lower capital costs [94].

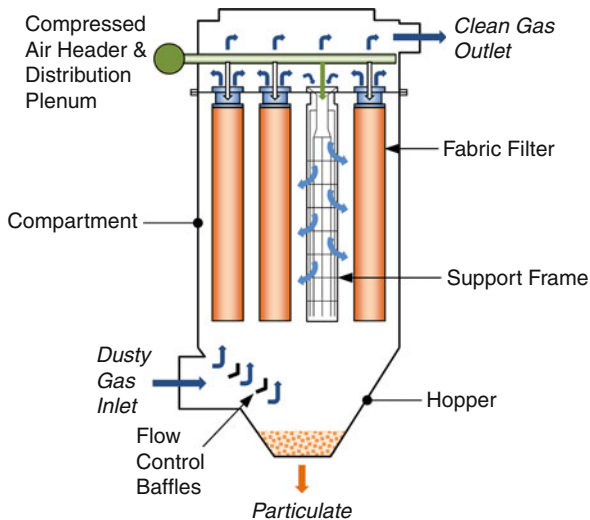
The filtration media used in fabric filters can be woven or felted. Woven fabrics are typically used in shaker-type baghouses as they have good mechanical strength. Felted fabrics tend to be used for reverse air and pulse-jet fabric filters [95]. For proper operation, the type of fabric must be matched to the specific application. The gas stream temperature and chemical composition are critical factors that influence the fabric selection. For coal-fired power plants, the baghouse is installed after the air preheater, where normal gas temperatures are in the range of 120–175°C (250–350°F) and where the presence of sulfur oxides and moisture increases the potential for corrosion. In this application, the utility industry has primarily used coated fiberglass fabrics for reverse airflow baghouse designs as these fabrics are resistant to chemical attack and can withstand temperatures up to 260°C [96]. The primary coating in use today is polytetrafluoroethylene (PTFE), which improves bag lifetime and provides resistance to acidic attack and thermal excursions. For pulse-jet baghouses equipped units fired with low-sulfur coal, the bags are typically made of polyphenylene sulfide (PPS) [97].

The performance of a fabric filter is impacted by the particle properties, fabric properties, operating characteristics, cleaning method, and interrelationships



Pulverized Coal-Fired Boilers and Pollution Control. Figure 4

Reverse air flow baghouse features



Pulverized Coal-Fired Boilers and Pollution Control. Figure 5

Figure 5

Pulse jet baghouse features

between these parameters [98]. While the fundamental mechanisms involved in removing particles from the gas are understood, the highly complex nature of the collection process and the participation of the dust layer in the collection process make the design of fabric

filters more empirical in nature. The two primary design and operating parameters are the air-to-cloth ratio, which sets the velocity ratio passing through the filter and, hence, the relative size of the unit, and the pressure drop, which sets the energy consumption requirements [99]. These two parameters are related as baghouses with a small A/C ratio will have a lower pressure drop than baghouses with a higher A/C ratio, resulting in a trade-off between capital and operating costs. The bag cleaning method influences the A/C ratio and pressure drop as more energetic cleaning methods can reduce the A/C ratio (and unit size) for a given pressure drop [100].

Fabric filters are highly efficient collection devices for both coarse and fine particulate, with typical efficiencies of 99–99.9%. Due to the collection mechanisms involved in fabric filtration, particle removal performance is not as sensitive to particle size for these devices as with ESPs. In comparing the costs of fabric filters to ESPs, fabric filters become more economic where highly efficient removal of submicron particles are required [101]. As noted above, the design A/C ratio and operating pressure drop impact fabric filter performance. Fabric filter performance is also impacted by flue gas and particle characteristics.

The flue gas temperature impacts the volume of the flue gas and, hence, the operating A/C ratio, while the flue gas moisture content and composition can impact the characteristics of the dust cake. The particle size impacts the buildup (and removal) of the dust cake on the filter and, along with the particle composition, impacts the adhesion of particles to the filter surface and the cohesion of particles to each other. Reactions between the particles and constituents of the gas can also impact the cohesiveness of the dust cake. Finally, the cleaning cycle interval and intensity will also impact the overall performance of a fabric filter.

### SO<sub>2</sub> Emissions Control Technologies

The primary mechanisms for removal of a gaseous pollutant from a gas stream involve (1) absorption into a liquid, (2) gas/solid reaction, and (3) adsorption onto a solid. The first two mechanisms are the primary methods for removal of sulfur dioxide from coal-fired boiler flue gas. Commercial processes based upon these mechanisms can be categorized into whether the process is regenerable, where the sulfur compound is separated from the absorbent, or nonregenerable, where the sulfur compounds are thrown away with the absorbent [102]. The product from regenerable processes can be concentrated SO<sub>2</sub>, hydrogen sulfide, elemental sulfur, or sulfuric acid. SO<sub>2</sub> removal processes can be further characterized into wet processes, where the product is a solution or slurry, and dry processes, where a dry product is produced [103]. The majority of non-regenerable processes produce a throwaway waste; however, some wet processes can be modified to produce a gypsum by-product that can be sold if site-specific conditions permit.

The primary technologies for removing sulfur dioxide from coal-fired power plants are wet limestone (CaCO<sub>3</sub>) and dry lime (Ca(OH)<sub>2</sub>) scrubbing [104], typically referred to as wet and dry flue gas desulfurization (FGD). The preference for these technologies over regenerable processes is largely driven by the low cost and high availability of limestone and lime, by the relative simplicity of these processes in comparison to other tail end processes, and by the fact that power plants are not chemical companies. For wet scrubbing, limestone tends to be preferred over lime due to lower cost, particularly if a ready supply is located close to the

power plant. In comparison to lime, limestone requires finer grinding, has higher transportation costs, requires larger equipment, and is less responsive with respect to pH control [105]. Thus, absorbent selection is done by performing a comparison based upon site-specific technical and economic factors. For wet FGD systems that produce gypsum, which will be sold for wallboard manufacturing, the use of hydrated lime rather than limestone can lead to improvements in the product quality in some cases [106]. In dry scrubbing, lime is required due to the need for rapid reaction times as lime has a higher reactivity than limestone.

A wide variety of wet and dry FGD processes exist, but the basic processes are similar [107–109]. In addition to or instead of lime and limestone, several processes introduce sodium compounds, such as sodium carbonate (e.g., trona), where these materials are available locally to improve process efficiency [110]. Seawater scrubbing, which uses seawater to remove the SO<sub>2</sub> and discharge it into the ocean, has also been applied to coal-fired power plants located in a coastal environment [111]. While wet and dry scrubbing are highly efficient technologies, they have high capital and operating costs. This has led to interest in lower capital cost SO<sub>2</sub> removal technologies such as furnace sorbent injection [112, 113]. The remainder of this section will discuss the basic wet limestone and dry lime scrubbing processes.

### Wet Flue Gas Desulfurization

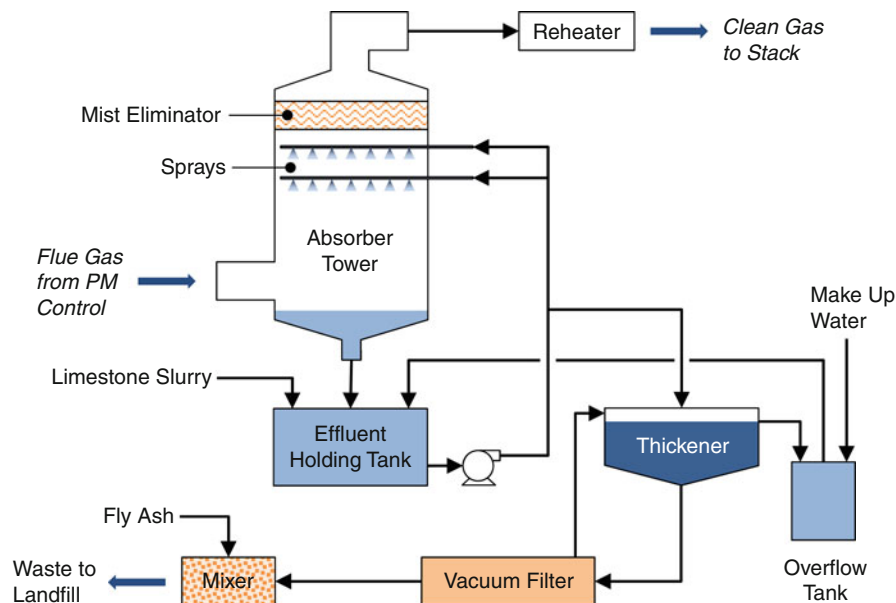
In wet FGD, flue gas containing SO<sub>2</sub> is input into a scrubbing tower or absorber where it is brought into contact with an alkaline solution or slurry containing partially dissolved limestone (CaCO<sub>3</sub>). The SO<sub>2</sub> is absorbed into the solution where it reacts to form calcium sulfite (CaSO<sub>3</sub>) and calcium sulfate (CaSO<sub>4</sub>). The detailed chemistry of the process is complex and depends upon the composition of the slurry and extent of dissolution of the limestone [105]. The primary steps involved in the process are gas/liquid mass transfer of the SO<sub>2</sub>, dissolution of the limestone, oxidation of SO<sub>3</sub> to SO<sub>4</sub> in solution, and crystallization of CaSO<sub>3</sub> and CaSO<sub>4</sub>. By controlling the solution pH and concentration of CaCO<sub>3</sub> in the solution, high levels of SO<sub>2</sub> removal can be achieved. As SO<sub>2</sub> absorbs into the solution as bisulfate (HSO<sub>3</sub><sup>-</sup>), the primary product

from the process is calcium sulfite. However, due to the high oxygen content of flue gas (3–10% by volume), a fraction of the calcium sulfite will oxidize to calcium sulfate. Typically, the product from this process is a waste sludge. It is possible to force oxidation of calcium sulfite into calcium sulfate by bubbling air through the solution which improves the ability to remove water from the waste and permits the material to be upgraded for use for gypsum ( $\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$ ).

The basic components of the limestone wet scrubbing process are illustrated in Fig. 6. Flue gas from the particulate control device (ESP or baghouse) enters the bottom of an absorber tower (typically a spray tower) and flows upward. The absorbing solution is sprayed into the gas stream. The  $\text{SO}_2$  is absorbed into the solution where it reacts with the limestone to form solids. The resulting slurry is then captured in an effluent holding tank where ground limestone slurry and recycled solution are added. This solution is then recycled to the absorber. A portion of the solution is sent to the thickener to concentrate the waste solids as a sludge. The sludge from the thickener is then dewatered in a vacuum filter to produce a filter cake, which is mixed with fly ash to stabilize the waste prior to sending it to a landfill. Enhancements to the basic

process include the use of a pre-scrubber upstream of the main absorber to quench the flue gas and protect the materials used in the absorber, incorporation of a second scrubbing loop into the pre-scrubber, and blowing air into the effluent holding tank to force oxidation of calcium sulfite to calcium sulfate [114]. This latter improvement minimizes scale formation, improves the characteristics of the sludge, and eliminates the need for adding fly ash for stabilization. Forced oxidation is also used to produce solids that can be used to produce gypsum for wallboard manufacturing. If lime rather than limestone is used, forced oxidation is not required for scale control. The cleaned flue gas exiting the absorber after passing through a mist eliminator designed to remove entrained droplets from the gas, passes through a reheater, and is then sent to the stack. A reheater is used to protect equipment downstream of the scrubber from condensation and corrosion, to reduce the visible plume, and to improve rise and dispersion of the stack gas.

The performance of a wet FGD system is impacted by a number of design and operational factors, most of which are related [114]. The design of the absorber tower must provide for effective contacting of the flue



**Pulverized Coal-Fired Boilers and Pollution Control. Figure 6**  
Basic limestone wet scrubbing process

gas with the absorber solution and for separation of the cleaned gases from the liquid. From an economic standpoint, the size of the tower is linked closely to the process chemistry and the relative flows of flue gas and absorber solution, which is typically expressed as the liquid-to-gas ratio. Operational factors that impact performance include the  $\text{SO}_2$  concentration in the flue gas, pH of the slurry, solids concentration in the slurry, concentration of other components in the slurry such as magnesium and chloride ions, residence time of the slurry in the reaction tank, and degree of slurry oxidation [105, 115]. High utilization of the limestone (or lime) is important and most systems operate with a calcium-to-sulfur molar ratio of close to 1:1. Liquid pH is monitored and used to control the limestone addition rate to maximum limestone utilization, to minimize the potential for scale in the system, and to reach a specific emissions target. The limestone grind fed to the FGD system can also be important, with finer grinds leading to a smaller-sized reaction tank, higher  $\text{SO}_2$  removals, and an increase in gypsum purity [116].

Wet FGD systems are capable of achieving over 95%  $\text{SO}_2$  removal, with some systems being designed for 96–98%, and can be applied to both low and high sulfur coals. As discussed above, a number of parameters can impact system performance and must be optimized to ensure effective  $\text{SO}_2$  control as well as maintain system reliability and availability. Overall, wet FGD systems represent a significant capital, operating, and maintenance expense for a coal-fired power plant. The main operating and maintenance (O&M) costs differ for various designs and should be considered in selecting an FGD process [117]. Experience with existing plants has led to a number of equipment improvements, such as large, single-tower, absorbers with higher flue gas velocities and improved mist eliminator designs, and process improvements, such as the use of fine limestone and solution buffers [103]. These improvements can lead to higher  $\text{SO}_2$  removal efficiencies, lower capital costs, and reduced O&M costs.

### Dry Flue Gas Desulfurization

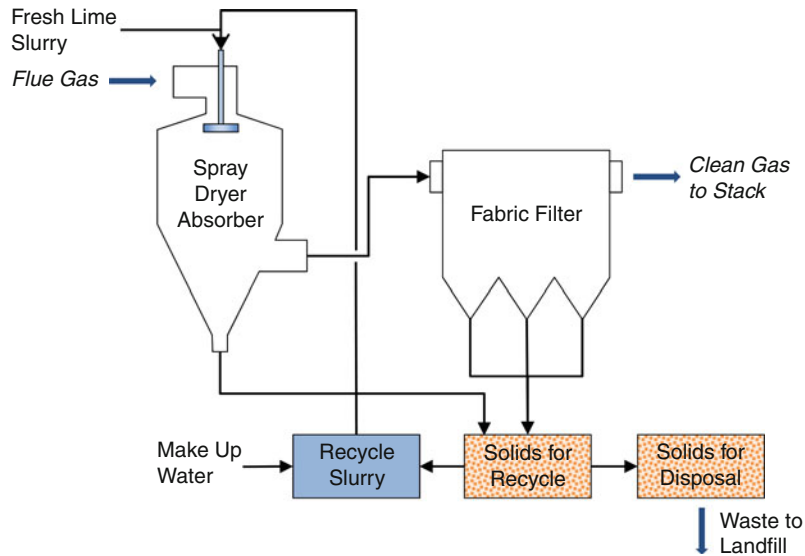
In dry FGD, flue gas containing  $\text{SO}_2$  is input into a spray dryer absorber where it is brought into contact with an alkaline solution containing hydrated lime ( $\text{Ca}(\text{OH})_2$ ). The  $\text{SO}_2$  is absorbed into the solution

where it reacts with the alkali material to form solid calcium sulfite ( $\text{CaSO}_3$ ). Absorption and reaction occur while the thermal energy of the flue gas vaporizes the water in the droplets to produce a fine powder. The primary processes involved in spray drying consist of atomization of the alkaline slurry into a spray of droplets, contacting of the spray with the flue gas, absorption of  $\text{SO}_2$  into the droplets, reaction with the suspended alkali material, drying of the spray by evaporation of the moisture in the droplets, and, finally, separation of the solids from the gas [118]. Due to the high oxygen concentrations of the flue gas, a fraction of the  $\text{CaSO}_3$  will be oxidized into  $\text{CaSO}_4$ .

An overview of the lime dry scrubbing process is shown in Fig. 7. Hot flue gas from the air preheater outlet enters the top of the cylindrical spray dryer where an atomizer sprays the alkaline slurry from the feed preparation system into the flue gas.  $\text{SO}_2$  is absorbed into the spray droplets where it reacts with the lime to form a solid. The water in the droplet evaporates leaving a fine powder that can be collected from the spray dryer and in a subsequent particulate collection device. Twin fluid and rotary or spinning disk atomizers can be used; however, most large utility applications use a rotary atomizer. Typically, fabric filters are used for removal of the fly ash and spent absorbent. Collection of the solids on the fabric filter permits additional  $\text{SO}_2$  removal to occur resulting in increased sorbent utilization.

The spray dryer absorber needs to be designed to maximize  $\text{SO}_2$  removal, yet produce a dry product. Since absorption and reaction of  $\text{SO}_2$  is fastest when surface liquid is present, it is necessary to operate as close to the saturation point of the flue gas as possible and to maximize the residence time available for reaction while still yielding dry solids at the absorber walls and outlet [119]. The quality of water that can be used for the slurry is limited by the boiler outlet temperature, which is set by the boiler thermal efficiency, and the desired approach to saturation. The concentration of solids in the slurry is limited by the viscosity of the material which impacts the practical aspects of handling and atomizing the material. These two factors constrain the  $\text{SO}_2$  removal that can be achieved for a given inlet  $\text{SO}_2$ .

Dry FGD systems are capable of achieving over 95%  $\text{SO}_2$  removal and are typically applicable to low and



**Pulverized Coal-Fired Boilers and Pollution Control. Figure 7**  
Basic lime dry scrubbing process

medium sulfur coals. There are technical constraints on the application of this technology to high sulfur coals as an increase in sulfur content requires an increase in the solids' content of the slurry and a point can be reached where the viscosity of the material is too high for proper pumping and atomization [103].

### NO<sub>x</sub> Emissions Control Technologies

Technologies to reduce NO<sub>x</sub> emissions from combustion systems can be placed into two broad categories: (1) combustion modification technologies which reduce the formation of NO<sub>x</sub> during combustion and (2) postcombustion technologies which eliminate NO<sub>x</sub> from the flue gas following combustion [28]. Combustion modification techniques will generally seek to lower flame temperature or to control the flame stoichiometric ratio depending upon whether thermal NO<sub>x</sub> or fuel NO<sub>x</sub> emissions are being targeted. For pulverized coal combustion, as discussed previously, the majority of the NO<sub>x</sub> emissions result from the nitrogen in the fuel; hence, the most effective combustion modifications are those which control the stoichiometric ratio during coal devolatilization. Postcombustion techniques tend to rely on the use of additives to reduce the NO<sub>x</sub> to molecular nitrogen, N<sub>2</sub>.

Removal of NO<sub>x</sub> by wet scrubbing is difficult due to the insolubility of NO and poor solubility of NO<sub>2</sub>. Scrubbing technologies that can be applied to industrial plants have not been widely applied to large pulverized coal-fired power plants due to high capital and operating costs and the low initial NO<sub>x</sub> concentrations [120].

For coal-fired power plants, the primary combustion modification techniques are low-NO<sub>x</sub> burners, overfire air, and reburning technology [121]. Each of these technologies relies on controlling the combustion process in the furnace. The primary postcombustion technologies are selective noncatalytic reduction (SNCR) and selective catalytic reduction (SCR) [122]. These technologies involve the use of a reagent that reduced NO<sub>x</sub> to N<sub>2</sub>. Combustion modification technologies are generally the most cost-effective when evaluated against postcombustion technologies on a cost per mass of pollutant removed [123]. SCR technology is the most expensive technology but also provides the highest level of NO<sub>x</sub> emission reduction. SNCR technology provides a modest level of NO<sub>x</sub> reduction, with cost-effectiveness typically between that of combustion modifications and SCR.

In retrofitting existing units with NO<sub>x</sub> emissions control technologies, selection of the most cost-effective technology or set of technologies generally



requires a boiler-specific analysis that includes assessment of the  $\text{NO}_x$  emission goals or limits and the expected performance and costs of the specific technologies. Retrofit costs for  $\text{NO}_x$  control technologies are very site specific and depend upon the ease of the retrofit and the boiler design and operating characteristics, such as the furnace size, firing configuration, and the condition of existing equipment [124]. The remainder of this section will provide an introduction to  $\text{NO}_x$  control technologies.

### Low- $\text{NO}_x$ Burners

Low- $\text{NO}_x$  burners reduce  $\text{NO}_x$  formation in the combustion process by delaying the mixing of fuel and air in the flame [125]. One means of aerodynamically staging fuel and air mixing in a coal flame for a wall-fired boiler is illustrated in Fig. 8. In this low- $\text{NO}_x$  burner design, the mixing of fuel and air is controlled by dividing the combustion air into multiple streams, with separate control over each stream. Coal entering the flame is initially burned in a fuel-rich core which gradually becomes fuel lean as air is progressively mixed into the flame. Staging of the air into the flame permits volatile nitrogen compounds released in the early portion of the flame to be processed in a reducing environment, resulting in lower  $\text{NO}_x$  emissions. Staging of the flame also reduces peak flame temperatures and, therefore, can reduce thermal  $\text{NO}_x$  formation as well.

For wall-fired boilers, low- $\text{NO}_x$  burner designs vary from manufacturer to manufacturer [126–128], but all tend to have similar features, including separate control over the combustion air split into two or more passages, control over the degree of swirl imparted to the

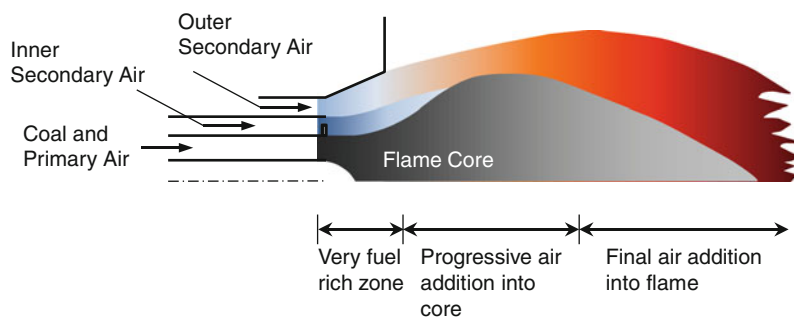
air streams, and a flame stabilizing device. The inner and outer air split and swirl settings are adjusted to generate a stable low- $\text{NO}_x$  emissions flame for a specific coal. These settings can also be adjusted to ensure that the flames do not impinge on the side or rear walls of the boiler.

For tangentially fired boilers, low- $\text{NO}_x$  burner designs typically consist of alternating passages of air and fuel [129]. The fuel nozzles direct the fuel into the center of the furnace, while the air nozzles direct the air closer to the furnace walls. Offsetting the air from the fuel generates a fuel-rich flame in the center of the furnace which leads to low- $\text{NO}_x$  emissions. The offset air also provides a measure of protection along the wall from rich conditions that can lead to wall corrosion. Additional air staging is achieved by adding a fraction of the combustion air through air ports located on the top of the burner stack. The air added to the top of the burner is typically called “close-coupled overfire air” or CCOFA.

The performance of low- $\text{NO}_x$  burners in a particular application is very dependent upon the boiler and coal characteristics. Low- $\text{NO}_x$  burners generally produce longer flames than the original equipment; hence, the trade-off between optimal burner performance and acceptable carbon-in-ash levels is strongly tied to the furnace geometry.  $\text{NO}_x$  control levels for the application of low- $\text{NO}_x$  burners to wall-fired utility boilers typically range from 40% to 50%.

### Overfire Air

Overfire air is a combustion modification technology that, like low- $\text{NO}_x$  burners, stages the combustion process to reduce  $\text{NO}$  emissions [130]. To apply



**Pulverized Coal-Fired Boilers and Pollution Control. Figure 8**  
Typical low- $\text{NO}_x$  burner design approach

overfire air to a coal-fired boiler, combustion air is diverted from the main combustion zone and is injected through ports located on the walls above the burners. In this process, the primary zone is operated slightly less fuel lean than usual, and fuel and air mixing is delayed. This delay reduces the formation of fuel NO and also reduces peak flame temperatures, resulting in a reduction in thermal NO formation as well. Overfire air is added to complete combustion of unburned fuel.

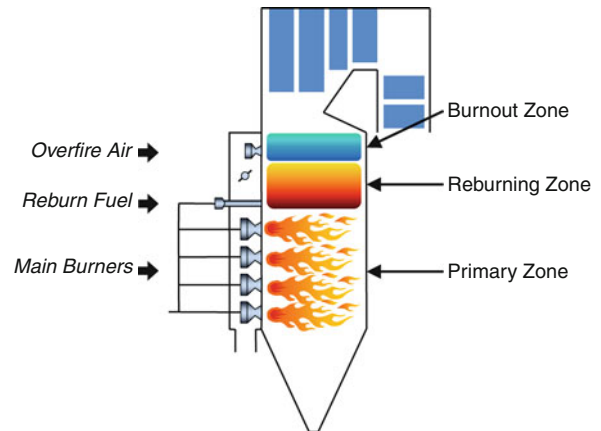
For pulverized coal-fired boilers, the performance of overfire air is dependent upon the overfire air system design and the burner/boiler characteristics. Optimal NO<sub>x</sub> control for a particular system can be achieved provided that the overfire air jets mix effectively, that good control over the main flame can be maintained as combustion air is diverted to the overfire air system, and that there is sufficient residence time in the furnace for carbon burnout to occur. As with low-NO<sub>x</sub> burners, the NO<sub>x</sub> control performance of overfire air is limited by acceptable carbon-in-ash. For pulverized coal-fired industrial boilers, NO<sub>x</sub> reductions between 15% and 30% are generally expected with the application of overfire air.

### Reburning

Reburning is a combustion modification technology that removes NO<sub>x</sub> from combustion products by using fuel as the reducing agent [131]. As illustrated in Fig. 9, application of reburning to a coal-fired boiler conceptually divides the furnace into three zones (primary zone, reburning zone, and burnout zone) which are typically defined as follows:

**Primary Zone:** The primary zone consists of the normal firing system and constitutes the bulk (typically 80–90%) of the total heat release. This zone is operated fuel lean at an excess air level which is close or slightly below normal operation. NO<sub>x</sub> formed in this zone enters the reburning zone.

**Reburning Zone:** The reburning fuel is injected downstream of the primary zone to create a slightly fuel-rich, NO<sub>x</sub> reduction zone. The reburning fuel provides the remainder, normally 10–20%, of the total heat input. In this zone, hydrocarbon radicals generated during breakdown of the reburning fuel react with NO molecules to form other reactive nitrogenous species, such as HCN,



### Pulverized Coal-Fired Boilers and Pollution Control.

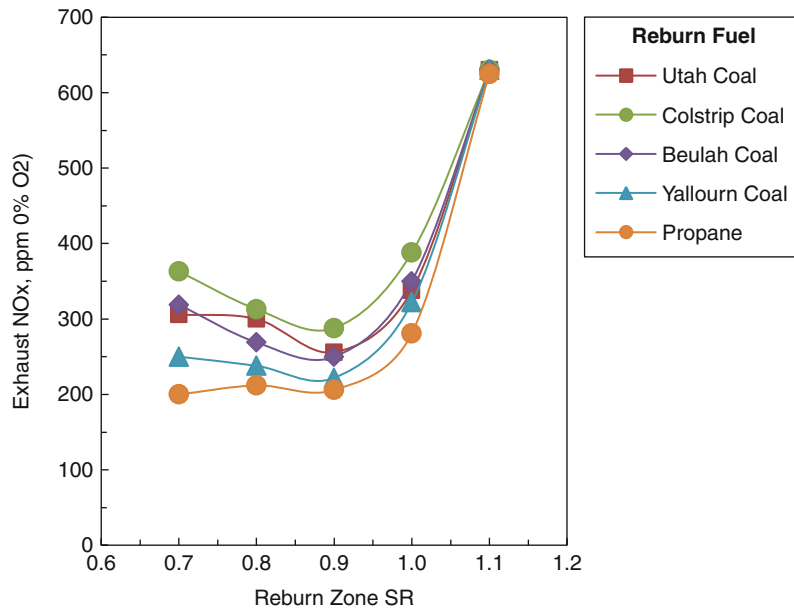
**Figure 9**

Application of reburning to a utility boiler

which react with other NO molecules to form molecular nitrogen, N<sub>2</sub>.

**Burnout Zone:** In this final zone, combustion air is added to complete the oxidation of the reburning fuel and to bring the boiler back to a normal operating excess air level.

Studies of the reburning process have shown that natural gas, fuel oil, and coal can be used as reburning fuels [132]. As shown in Fig. 10, hydrocarbon fuels with low fuel-bound nitrogen, such as propane or natural gas, provide the lowest NO<sub>x</sub> emissions since nitrogen in the reburning fuel tends to contribute to the final NO<sub>x</sub> exiting the process. In Fig. 10, the reburning zone stoichiometric ratio is defined as the ratio of the total air supplied to the primary and reburning zones to the total stoichiometric air requirements of the primary and reburning fuels. The parameters which control the performance of the reburning process have been defined through an extensive series of experimental studies [133]. For boiler applications, the most critical parameters are the reburning zone stoichiometric ratio and the initial NO<sub>x</sub> level. As shown in Fig. 10, most fuels exhibit a maximum NO<sub>x</sub> reduction (i.e., minimum NO<sub>x</sub> emissions) at a reburning stoichiometric ratio of about 0.9, which corresponds to a reburning fuel input of approximately 18–19% of the total heat input at a primary zone excess air level of 10%. NO<sub>x</sub> reduction with reburning increases as the NO<sub>x</sub> level from the primary zone increases.



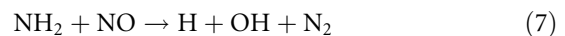
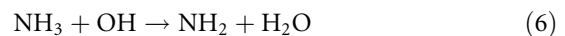
**Pulverized Coal-Fired Boilers and Pollution Control. Figure 10**  
Impacts of reburning zone stoichiometry and fuel on reburning performance

Similar to other combustion modification techniques, the most significant factors which limit the performance of reburning in full-scale applications are the design of the reburning system and the boiler characteristics. The boiler geometry must permit adequate mixing of the reburning fuel and overfire air to be achieved and must provide adequate residence time for the process to be implemented above the primary combustion zone. Reliable methodologies for scaling the reburning process to boilers with varying characteristics have been developed [134]. The use of these methodologies has led to  $\text{NO}_x$  control levels between 50% and 70% for application of reburning to full-scale utility boilers, independent of the firing configuration.

### Selective Noncatalytic Reduction

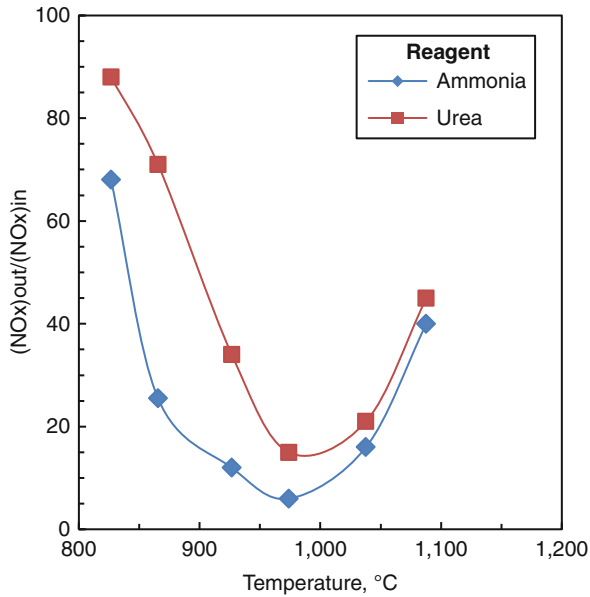
Selective noncatalytic reduction is a flue gas treatment process in which an amine-containing agent, such as ammonia ( $\text{NH}_3$ ) or urea ( $\text{CO}(\text{NH}_2)_2$ ), is injected into combustion gases to react with and reduce NO formed during the combustion process [135, 136]. At the proper temperature,  $\text{NH}_2$ , generated from decomposition of the injected reagent, reacts with NO via a complex set of reactions to form  $\text{N}_2$ . The chemistry

of the SNCR process is illustrated in the following global reactions that are based upon the use of ammonia as the reagent:



Initially, the injected ammonia is activated by OH to form a reactive amine,  $\text{NH}_2$ , radical. This radical then selectively reacts with NO to form molecular nitrogen,  $\text{N}_2$ , and H and OH radicals. The OH radical can help to activate additional ammonia or can react with H radicals to form water,  $\text{H}_2\text{O}$ .

Figure 11 shows the impacts of gas temperature on the  $\text{NO}_x$  reduction achieved with urea or ammonia in tests performed in a pilot-scale combustion facility. As shown in this figure, at the proper flue gas temperature (approximately 950–1,000°C), the injected reagent selectively reduces NO to molecular nitrogen,  $\text{N}_2$ . As temperatures increase,  $\text{NO}_x$  reduction performance is reduced and a portion of the reagent can be oxidized to NO. As temperatures decrease, a portion of the reagent exits the process unconverted. The unconverted reagent exiting the process is referred to as ammonia slip.



**Pulverized Coal-Fired Boilers and Pollution Control.**

**Figure 11**

Impacts of gas temperature in SNCR performance

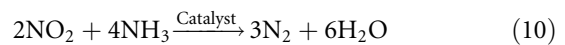
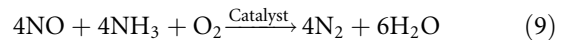
The performance of SNCR on utility boilers depends upon the reagent used, the quantity of reagent injected, and the boiler design and operating characteristics [137]. For maximum performance, the injection system must be designed to provide effective mixing of the reagent within the optimal temperature window. In addition, the boiler design should provide adequate residence time at the proper temperatures. For pulverized coal-fired boilers, the optimal temperature window generally resides within the convective heat transfer sections, where access is limited and quench rates are high [138]. Wide variations in temperature at the point of injection due to stratification or to changes in load can also limit performance. Multiple injection locations can help to reduce the negative impacts of temperature variations. Due to concerns over the handling and storage of anhydrous ammonia, aqueous ammonia or urea is generally preferred for application of SNCR to utility boilers.

Application of SNCR to coal-fired boilers is expected to reduce NO<sub>x</sub> emissions by 30–60% from uncontrolled levels. The higher levels of reduction are generally achieved in boilers with slower quench rates

(temperature decay versus residence time) in the upper furnace and with good access to the optimal temperature window.

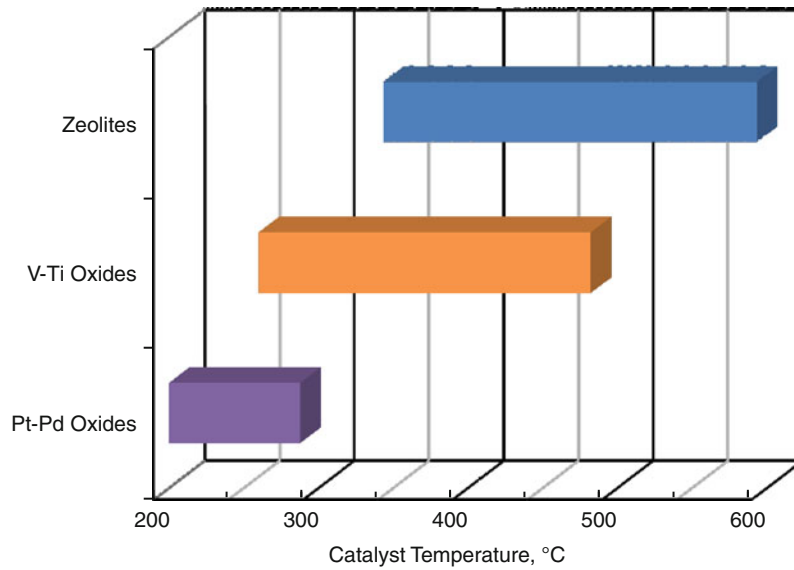
### Selective Catalytic Reduction

Selective catalytic reduction is a high-efficiency flue gas treatment process in which ammonia is added to the flue gas to react with NO over a catalyst [139]. The primary overall reactions for this process can be expressed as:



SCR catalysts are made of a solid ceramic material that contains active catalytic components. The ceramic may be in the form of a honeycomb monolith or it may be wash coated onto a ceramic, metal, or fiber substrate [140]. The typical operating temperature range for available catalysts is shown in Fig. 12. Noble metal catalysts, such as platinum-palladium formulations, function at lower temperatures (150–290°C; 300–550°F) than base metal and zeolite catalysts but are susceptible to deactivation by sulfur oxides. Base metal catalysts, such as vanadium-titanium formulations, function at temperatures in the range of 260–490°C (500–900°F) and are more resistant to sulfur deactivation. Zeolite catalysts operate at a temperature range above that of the other catalysts (340–590°C, 650–1,100°F) and have shown a good resistance to poisoning from trace elements in coal such as arsenic. The majority of SCR technology applications to pulverized coal-fired boilers use vanadia-titania catalysts as they are widely available and have good resistance to deactivation by sulfur oxides in the flue gas. Catalyst formulations are specific to the catalyst manufacturer and may include the addition of other base metal oxides to increase the catalyst activity, to increase poison resistance, to minimize ammonia slip, and to reduce SO<sub>2</sub> oxidation to SO<sub>3</sub>.

To meet the temperature requirements for vanadia-titania catalysts, the catalyst needs to be installed in the flue gas path between the boiler economizer and air preheater. The arrangement for both new and retrofit SCR systems depends upon the boiler and balance of plant equipment arrangement [141, 142]. Typically,



**Pulverized Coal-Fired Boilers and Pollution Control. Figure 12**  
Operating range for various SCR catalyst formulations

one or two layers of catalyst are incorporated into a catalyst housing or reactor. Flue gas from the economizer exit is either routed in a horizontal duct to the top of the reactor where it then flows downward over the catalyst. Alternatively, the ductwork arrangement may need a vertical upward duct followed by a 180° transition into the top of the reactor. Ammonia is injected into the flue gas upstream of the SCR reactor using an ammonia injection grid that typically consists of a series of injection lances spaced across the duct and designed to provide good mixing of the ammonia with the flue gas. Static mixers and flow conditioning devices such as turning vanes are used to ensure uniform distribution of the ammonia and flue gas within the ductwork.

A number of factors need to be considered in designing an SCR system, including reactor location and design, catalyst type and configuration, and the location and design of the ammonia injection grid. The catalyst space velocity and uniformity of ammonia and flue gas distribution all influence system performance. Coal properties are an important consideration for catalyst selection [143]. In general, since the SCR system is separate from the combustion process and since the extensive retrofit requirements permit the design to be optimized, the overall control performance is less influenced by boiler characteristics than the other NO<sub>x</sub>

control technologies described previously. On the other hand, the cost of installing an SCR system is very sensitive to site-specific factors due to the extensive retrofit requirements [144]. NO<sub>x</sub> reduction in the range of 70–90% is expected for application of SCR to coal-fired boilers.

### Future Directions

Significant steps have been taken to reduce air pollutant emissions from coal-fired power plants throughout the developed countries in the world. Promulgation of more stringent emission limits in various counties will require further application of these technologies to smaller power plants and place constraints on new power plants. The cost of the air pollution control technologies is a challenge with both retrofit and new technologies and, hence, research and technology development is focused on reducing both the capital and operating costs of these technologies, on improving performance to meet stricter standards, and on improving equipment reliability and availability. In addition, due to concerns over emissions of air toxic compounds from coal-fired power plants, the potential positive and negative impacts of existing air pollution control equipment on these emissions is an area of active investigation.

## Bibliography

1. International Energy Agency (2010) World energy outlook 2010. International Energy Agency, Paris
2. World Coal Institute (2009) The coal resource, a comprehensive overview of coal. World Coal Institute, London
3. Freese B (2003) Coal: a human history. Perseus, Cambridge, MA
4. Stuart R (1824) A descriptive history of the steam engine. John Knight and Henry Lacey, London
5. Burn RS (1854) The steam engine. H. Ingram, London
6. Thurston RH (1903) A history of the growth of the steam engine, 4th edn. D. Appelton, New York, Revised
7. Tagliaferro L (2003) Thomas Edison: inventor of the age of electricity. Lerner, Minneapolis
8. Woodside M (2007) Thomas A. Edison: the man who lit up the world. Sterling, New York
9. Babcock & Wilcox Company (1992) Steam: its generation and use, Fortieth edition. Babcock & Wilcox, Barberton
10. Combustion Engineering, Inc (1981) Combustion: fossil power systems, 3rd edn. Combustion Engineering, Windsor
11. Heidorn KC (1978) A chronology of important events in the history of air pollution meteorology to 1970. Bulletin of the American Meteorological Society, vol 59, issue 12. American Meteorological Society, Toronto, pp 1589–1597
12. Thorshiem P (2006) Inventing pollution: coal, smoke, and culture in Britain since 1800. Ohio University Press, Athens
13. Jacobs C, Kelly WJ (2008) Smogtown: the lung-burning history of pollution in Los Angeles. The Overlook Press, New York
14. Brimblecombe P (1987) The big smoke: a history of air pollution in London since medieval times. Methuen, London
15. Greater London Authority (2002) 50 years on: the struggle for air quality in London since the great smog of December 1952. Greater London Authority, London
16. Davis DL (2002) When smoke ran like water: tales of environmental deception and the battle against pollution. Basic Books, New York
17. Reitze AW (2005) Stationary source air pollution law. Environmental Law Institute, Washington, DC
18. Martineau RJ, Novello DP (2004) The clean air act handbook. American Bar Association, Chicago
19. U.S. Environmental Protection Agency (2005) Review of the National Ambient Air Quality Standards for particulate matter. Policy assessment of scientific and technical information. OAQPS staff paper. EPA-425/R-05-005a. Office of Air Quality Planning and Standards, Research Triangle Park
20. U.S. Environmental Protection Agency (2009) Integrated science assessment for particulate matter. Office of Research and Development, Research Triangle Park. EPA/600/R-08/139F
21. U.S. Environmental Protection Agency (2004) Air quality criteria for particulate matter. EPA/600/P-99/002aF. Office of Research and Development, Research Triangle Park
22. U.S. Environmental Protection Agency (2008) Integrated science assessment for sulfur oxides – health criteria. EPA/600/R-08/047F. Office of Research and Development, Research Triangle Park
23. U.S. Environmental Protection Agency (2008) Integrated science assessment for oxides of nitrogen – health criteria. EPA/600/R-08/071. Office of Research and Development, Research Triangle Park
24. Glass NR (1979) Environmental effects of increased coal utilization: ecological effects of gaseous emissions from coal combustion, vol 33. U.S. National Institute of Environmental Health Sciences, Research Triangle Park, pp 249–272
25. Bennett DA, Goble RL, Linhurst RA (1985) The acidic deposition phenomenon and its effects: critical assessment document. EPA/600/8-85/001. U.S. Environmental Protection Agency, Office of Research and Development, Research Triangle Park
26. National Acid Precipitation Assessment Program (2005) National Acid Precipitation Assessment Program report to Congress: an integrated assessment. National Acid Precipitation Assessment Program
27. U.S. Environmental Protection Agency (1997) Nitrogen oxides: impacts on Public Health and the Environment. EPA-452-R-97-002. Office of Air and Radiation, Research Triangle Park
28. U.S. Environmental Protection Agency (1999) Nitrogen oxides (NOx), why and how they are controlled. EPA-456/F-99-006R. Office of Air Quality Planning and Standards, Research Triangle Park
29. McConville A (1997) Emissions standards handbook. IEACR/96. IEA Coal Research, London
30. U.S. Environmental Protection Agency (2000) Air quality criteria for carbon monoxide. EPA 600/P-99/001F. Office of Research and Development, Washington, DC
31. U.S. Environmental Protection Agency (1998) Study of hazardous air pollutant emissions from Electric Utility Steam Generating Units. Final report to congress, vols 1 and 2. EPA-453/R-98-004a, b. Office of Air Quality Planning and Standards, Research Triangle Park
32. U.S. Environmental Protection Agency (1997) Mercury study report to congress, vols I–VII. EPA-452/R-97-003, 010. Office of Air Quality Planning and Standards, Research Triangle Park
33. U.S. Environmental Protection Agency (2000) Regulatory finding on the emissions of hazardous air pollutants from Electric Utility Steam Generating Units, vol 65(245). Federal Register, Research Triangle Park, pp 79825–79831
34. National Research Council (2000) Toxicological effects of methylmercury. National Academy Press, Washington, DC
35. European Commission (2001) Ambient air pollution by mercury (Hg). Position paper. Office for Official Publications of the European Communities, Luxembourg
36. Environment Canada (2010) Risk management strategy for mercury
37. UNEP Chemicals Branch, DTIE (2010) The global atmospheric mercury assessment: sources, emissions and transport. United Nations Environment Programme, Geneva
38. US Environmental Protection Agency (2011) National Emission Standards for Hazardous Air Pollutants from Coal and Oil-Fired Electric Utility Steam Generating Units and Standards of

- Performance for Fossil Fuel-Fired Electric Utility, Industrial-Commercial- Institutional, and Small Industrial-Commercial-Institutional Steam Generating Units, vol 76(85). Federal Register, Research Triangle Park, pp 24976–25147
39. Mitchell JFB (1989) The “Greenhouse” effect and climate change, vol 27, No 1. American Geophysical Union, Washington, DC, pp 115–139
  40. U.S. Environmental Protection Agency (1995) Compliance of air pollutant emissions factors, 5th edn. AP-42. Office of Air Quality Planning and Standards, Research Triangle Park
  41. Eggleston HS, Buendia L, Miwa K, Ngara T, Tanabe K (eds) (2006) 2006 IPCG guidelines for National Greenhouse Gas Inventories. National Greenhouse Gas Inventories Programme. Intergovernmental Panel on Climate Change, Hayama, Kanagawa
  42. Ciferno JD, Fout TE, Jones AP, Murphy JT (2009) Capture carbon from existing coal-fired power plants, *Chemical Engineering Progress*. American Institute of Chemical Engineers, New York, pp 33–41
  43. IRA Greenhouse R&D Programme (2007) CO<sub>2</sub> capture ready plants. Report No. 2007/4. International Energy Agency, Cheltenham
  44. Newton GH, Schieber C, Socha RG, Kramlich JC (1990) Mechanisms governing fine particulate emissions from coal flames. DOE/PC/73743-T8. U.S. Department of Energy, Pittsburgh
  45. Gluskoter HJ (1978) An introduction to the occurrence of mineral matter in Coal. [ed.] Richard W. Bryers. Ash deposits and corrosion due to impurities in combustion gases. Hemisphere Publishing Department, Washington, DC, pp 3–19
  46. Chen Y, Shah N, Huggins FE, Huffman GP, Linak WP, Andrew Miller C (2004) Investigation of primary fine particulate matter from coal combustion by computer-controlled scanning electron microscopy, vol 85. *Fuel Processing Technology*, Elsevier, Amsterdam, pp 743–761
  47. Folsom BA, Heap MP, Pohl JH (1986) State-of-the-art review summary and program plan. Effects of coal quality on power plant performance and costs, vol 1. Report No. CS-4283. Electric Power Research Institute, Palo Alto
  48. Miller CA, Linak WP (2002) Primary particles generated by the combustion of heavy fuel oil and coal. Review of research results from EPA’s National Risk Management Research Laboratory. EPA-600/R-02-093. U.S. Environmental Protection Agency, Office of Research and Development, Research Triangle Park
  49. Flagan RC, Seinfeld JH (1988) Fundamentals of air pollution engineering. Prentice Hall, Englewood Cliffs
  50. Flagan RC (1978) Submicron particles from coal combustion. In: Proceedings of the seventeenth symposium (international) on combustion. The Combustion Institute, Pittsburgh, pp 97–104
  51. Kramlich JC, Chenevert B, Park J, Hoffman DA, Butcher EK (1996) Suppression of fine ash formation in pulverized coal flames. Final Technical Report prepared for DOE Grant No. DE-FG22-92PC92548. U.S. Department of Energy, Pittsburgh Energy Technology Center, Pittsburgh
  52. Acurex Environmental Corporation (1993) Emission factor documentation for AP-42 Section 1.1 bituminous and subbituminous coal combustion. U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park
  53. Attar A, Dupuis F (1979) Data on the distribution of organic sulfur functional groups in coal. Preprints of the Papers of the Spring National Meeting of the Division of Fuel Chemistry, vol 24(1). American Chemical Society, Honolulu, pp 166–177
  54. Pershing DW, Silcox GD (1988) SO<sub>x</sub> fundamentals. Combustion of solid fuels – a one week intensive course. International Flame Research Foundation, Noordwijkerhout
  55. Davis K, Dissel A, Valentine J (2001) The evolution of pyritic and organic sulfur from pulverized coal particles during combustion. The 2nd joint meeting of the US Sections of The Combustion Institute. The Combustion Institute, Oakland
  56. Muller III CH, Schofield K, Steinberg M, Broida HP (1978) Sulfur chemistry in flames. In: Proceedings of the seventeenth symposium (international) on combustion. The Combustion Institute, Pittsburgh, pp 867–879
  57. Kramlich JC, Malte PC, Grosshandler WL (1981) The reaction of fuel-sulfur in hydrocarbon combustion. In: Proceedings of the eighteenth symposium (international) on combustion. The Combustion Institute, Pittsburgh, pp 151–161
  58. Shao D, Hutchinson EJ, Heidbrink J, Pan W-P, Chou C-L (1994) Behavior of sulfur during coal pyrolysis. *J Anal Appl Pyrolysis* 30:91–100. Elsevier, Amsterdam
  59. Castaldini C, Angwin M (1977) Boiler design and operating variables affecting uncontrolled sulfur emissions from pulverized-coal-fired steam generators. EPA-450/3-77-047. U.S. Environmental Protection Agency, Research Triangle Park
  60. Folkedahl BC, Zygarlicke CJ (2004) Sulfur retention in North Dakota lignite coal ash. Preprints of the papers of the spring national meeting of the Division of Fuel Chemistry, vol 49(1). American Chemical Society, Anaheim, pp 167–168
  61. Blythe G, Dombrowski K (2004) SO<sub>3</sub> mitigation guide update. Report No. 1004168. Electric Power Research Institute, Palo Alto
  62. Srivastava RK, Miller CA, Erickson C, Jambhekar R (2004) Emissions of sulfur trioxide from coal-fired power plants. *J Air Waste Manag Assoc* 54:750–762. Air and Waste Management Association, Pittsburgh
  63. Glassman I (1987) *Combustion*, 2nd edn. Academic, Orlando
  64. Fenimore CP (1971) Formation of nitric oxide in premixed hydrocarbon flames. In: Proceedings of the thirteenth symposium (international) on combustion. The Combustion Institute, Pittsburgh, pp 373–389
  65. Pershing DW, Wendt JOL (1977) The influence of flame temperature and coal composition on thermal and fuel NO<sub>x</sub>. In: Proceedings of the sixteenth symposium (international) on combustion. The Combustion Institute, Pittsburgh, pp 389–399
  66. Sarofim AF, Beér JM (1990) The fate of fuel nitrogen and ash during combustion of pulverized coal, Chapter 4. In: Lemieux PM, Air and Energy Engineering Research Laboratory (eds)

- Pulverized coal combustion: pollutant formation and control, 1970–1980. U.S. Environmental Protection Agency, Air and Energy Engineering Research Laboratory, Research Triangle Park
67. Hayhurst AN, Vince IM (1980) Nitric oxide formation from N<sub>2</sub> in flames. *Prog Energy Combust Sci* 6:35–51. Elsevier, Amsterdam
  68. Kelemen SR, Gorbaty ML, Vaughn SN, Kwiatek PJ (1993) Quantification of nitrogen forms in argonne premium coals. Preprints of the papers of the spring national meeting of the Division of Fuel Chemistry, vol 38(2). American Chemical Society, Denver, pp 384–392
  69. Mitra-Kirtley S., Mullins OC, Branthaver J, Van Elp J, Cramer SP (1993) Nitrogen XANES studies of fossil fuels. Preprints of the papers of the fall national meeting of the Division of Fuel Chemistry, vol 38(3). American Chemical Society, Chicago, pp 762–768
  70. Solomon PR, Fletcher TH (1994) Impact of coal pyrolysis on combustion. In: Proceedings of the twenty-fifth symposium (international) on combustion. The Combustion Institute, Pittsburgh, pp 463–474
  71. Pohl JH, Sarofim AF (1977) Devolatilization and oxidization of coal nitrogen. In: Proceedings of the sixteenth symposium (international) on combustion. The Combustion Institute, Pittsburgh, pp 491–501
  72. Rees DP, Smoot LD, Hedman PO (1981) Nitrogen oxide formation inside a laboratory pulverized coal combustor. In: Proceedings of the eighteenth symposium (international) on combustion. The Combustion Institute, Pittsburgh, pp 1305–1311
  73. Chen JC, Niksa S (1992) Suppressed nitrogen evolution from coal-derived soot and low-volatility coal chars. In: Proceedings of the twenty-fourth symposium (international) on combustion. The Combustion Institute, Pittsburgh, pp 1269–1276
  74. Zhang H (2001) Nitrogen evolution and soot formation during secondary coal pyrolysis. PhD dissertation, Brigham Young University
  75. Pohl JH, Chen SL, Heap MP, Pershing DW (1983) Correlation of NO<sub>x</sub> emissions with basic physical and chemical characteristics of coal. In: Proceedings of the 1982 joint symposium on NO<sub>x</sub> control, vol 2. Electric Power Research Institute, Palo Alto
  76. Pershing DW, Heap MP, Chen SL (1990) Bench-scale experiments on the formation and control of NO<sub>x</sub> emissions from pulverized coal combustion, Chapter 9. In: Pulverized coal combustion: pollutant formation and control, 1970–1980. U.S. Environmental Protection Agency, Air and Energy Engineering Research Laboratory, Research Triangle Park
  77. Chen SL, Heap MP, Pershing DW, Martin GB (1982) Influence of coal composition on the fate of volatile and char nitrogen during combustion. In: Proceedings of the nineteenth symposium (international) on combustion. The Combustion Institute, Pittsburgh, pp 1271–1280
  78. Bartok W, Crawford AR, Piegari GJ (1971) Systematic field study of NO<sub>x</sub> emission control methods for utility boilers. APTD-1163. U.S. Environmental Protection Agency, Office of Air Programs, Research Triangle Park
  79. Payne R, Heap MP, Pershing DW (1981) NO<sub>x</sub> formation and control in pulverized-coal flames. In: Proceedings of the low rank coal technology development workshop, 17–18 June 1981. San Antonio
  80. Licht W (1980) Air pollution control engineering, basic calculations for particulate collection. Marcel Dekker, New York
  81. Strauss W (1975) Industrial gas cleaning, 2nd edn. Pergamon Press, New York
  82. Kinsey JS, Schliesser S, Englehart PJ (1985) Control technology for sources of PM<sub>10</sub>. Draft report prepared for EPA Contract No. 68-02-03891, work assignment 4. U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park
  83. U.S. Environmental Protection Agency (1982) Control techniques for particulate emissions from stationary sources, vol 1. EPA-450/3-81-005a. Emissions Standards and Engineering Division, Research Triangle Park
  84. White HJ (1977) Electrostatic precipitation of fly ash. *J Air Pollut Control Assoc* 27(1):15–21. Air Pollution Control Association, Pittsburgh
  85. U.S. Environmental Protection Agency (2011) Electrostatic precipitator plan review. APTI Course No. SI:412B. U.S. Environmental Protection Agency, Research Triangle Park
  86. White HJ (1977) Electrostatic precipitation of fly ash. Fly ash and furnace gas characteristics. *J Air Pollut Control Assoc* 27(2):114–120. Air Pollution Control Association, Pittsburgh
  87. Turner JH, Lawless PA, Yamamoto T, Coy DW, Mckenna JD, Mycock JC, Nunn AB, Greiner GP, Vatauvuk WM (2002) Electrostatic precipitators. In: Mussatti DC (ed) EPA Air pollution control cost manual. U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park. Chapter 3 in Section 6, Particulate matter controls
  88. Szabo MF, Shah YM (1979) Inspection manual for evaluation of electrostatic precipitator performance. EPA-340/1-79-007. U.S. Environmental Protection Agency, Division of Stationary Source Enforcement, Washington, DC
  89. White HJ (1977) Electrostatic precipitation of fly ash. Precipitator design. *J Air Pollut Control Assoc* 27(3):206–217. Air Pollution Control Association, Pittsburgh
  90. National Air Pollution Control Administration (1969) Control techniques for particulate air pollutants. NAPCA Publication No. AP-51. National Air Pollution Control Administration, Washington, DC
  91. Turner JH, Lawless PA, Yamamoto T, Coy DW, Mckenna JD, Mycock JC, Nunn AB, Greiner GP, Vatauvuk WM (2002) Baghouses and filters. In: Mussatti DC (ed) EPA air pollution control cost manual. U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park. Chapter 1 in Section 6, Particulate matter controls
  92. Cushing KM, Merritt RL, Chang RL (1990) Operating history and current status of fabric filters in the utility industry. *J Air Waste Manage Assoc* 40(7):1051–1058. Air and Waste Management Association, Pittsburgh
  93. Belba VH, Theron Grubb W, Chang R (1992) The potential of pulse-jet baghouses for utility boilers. Part 1: a worldwide survey of users. *J Air Waste Manage Assoc* 42(2):209–217. Air and Waste Management Association, Pittsburgh



94. Sloat DG, Gaikwad RP, Chang RL (1993) The potential of pulse-jet baghouses for utility boilers. Part 3: comparative economics of pulse-jet baghouse precipitators and reverse-gas baghouses. *J Air Waste Manage Assoc* 43:120–128. Air and Waste Management Association, Pittsburgh
95. EC/R Incorporated (1998) Stationary source control techniques document for fine particulate matter. Report prepared for EPA Contract No. 68-D-98-026, Work Assignment No. 0-08. U.S. Environmental Protection Agency, Research Triangle Park
96. Felix LG, Cushing KM, Grubb WT, Giovanni DV (1988) Fabric filters for the electric utility industry. In: *Guidelines for fabrics and bags*, vol 3, CS-5161. Electric Power Research Institute, Palo Alto
97. Electric Power Research Institute (2005) Utility boiler baghouse update. Report No. 1010367. Electric Power Research Institute, Palo Alto
98. Dennis R (1974) Collection efficiency as a function of particle size, shape and density: theory and experience. *J Air Pollut Control Assoc* 23(12):1156–1161. Air Pollution Control Association, Pittsburgh
99. Carr RC, Smith WB (1974) Fabric filter technology for utility coal-fired power plants. *J Air Pollut Control Assoc* 34(1):79–89. Air Pollution Control Association, Pittsburgh
100. Carr RC, Cushing KM, Gallaer CA, Smith WB (1992) Fabric filters for the electric utility industry. In: *Guidelines for fabric filter design*, vol 5, CS-5161. Electric Power Research Institute, Palo Alto
101. Bustard CJ, Cushing KM, Pontius DH, Smith WB (1988) Fabric filters for the electric utility industry. In: *General concepts*, vol 1, CS-5161. Electric Power Research Institute, Palo Alto
102. Devitt T, Gestle R, Gibbs L, Hartman S, Klier R (1978) Flue gas desulfurization system capabilities for coal-fired steam generators, vol II. Technical report. U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC, EPA-600/7-78-032b
103. Wojciech J, Singer C, Srivastava RK, Tsigotis PE (1999) Status of SO<sub>2</sub> scrubbing technologies. EPRI-DOE-EPA combined utility air pollution control symposium. TR-113187-V1. Electric Power Research Institute, Palo Alto
104. Srivastava RK (2000) Controlling SO<sub>2</sub> emissions: a review of technologies. EPA/600/R-00/093. U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC
105. Kohl A, Riesenfeld F (1985) *Gas purification*, 4th edn. Gulf, Houston
106. Weiler H, Ellison W (1997) Wet gypsum-yielding FGD experience using quicklime reagent. EPRI-DOE-EPA combined utility air pollutant control symposium. The mega symposium. TR-108683-V2. Electric Power Research Institute, Palo Alto
107. United Engineers and Constructors, Inc (1991) Economic evaluation of flue gas desulfurization systems. GS-7193, vol 1. Electric Power Research Institute, Palo Alto
108. United Engineers and Constructors, Inc (1992) Economic evaluation of flue gas desulfurization systems. GS-7193, vol 2. Electric Power Research Institute, Palo Alto
109. United Engineers and Constructors, Inc (1995) Economic evaluation of flue gas desulfurization systems. GS-7193-V3, vol 3. Electric Power Research Institute, Palo Alto
110. Fox MR, Hunt TG (1990) Flue gas desulfurization using dry sodium injection. Presented at the EPA/EPRI 1990 SO<sub>2</sub> control symposium. s.n., New Orleans
111. Radojevic M (1991) Scrubbing of flue gases with sea-water. Presented at the AFRC/JFRC international conference on environmental control of combustion processes, 7–10 October 1991. s.n., Honolulu
112. Zhou W, Maly P, Brooks J, Nareddy S, Swanson L, Moyeda D (2010) Design and test furnace sorbent injection for SO<sub>2</sub> removal in a tangentially fired boiler. *Environmental Engineering Science*, vol 27, Number 4, Mary Ann Liebert, New Rochelle, pp 337–345
113. Nolan PS (1996) Emission control technologies for coal-fired power plants. Presented at the People's Republic of China Ministry of Electric Power seminar, 22–25 April 1996, s.n., Beijing
114. Henzel DS, Laseke BA, Smith EO, Swenson DO (1982) *Handbook for flue gas desulfurization scrubbing with limestone*. Noyes Data Corporation, Park Ridge
115. Miller SF, Miller BG, Scaroni AW (1997) Limestone performance in a pilot-scale forced oxidation scrubber. EPRI-DOE-EPA combined utility air pollutant control symposium. The mega symposium. TR-108683-V2, Electric Power Research Institute, Palo Alto
116. Brogen C, Klingspor JS (1997) Impact of limestone grind size on WFGD performance, August. EPRI-DOE-EPA combined utility air pollutant control symposium. The mega symposium. TR-108683-V2. Electric Power Research Institute, Palo Alto
117. Blythe G, Horton B, Rhudy R (1999) EPRI FGD operating and maintenance cost survey. EPRI-DOE-EPA combined utility air pollution control symposium. TR-113187-V1, Electric Power Research Institute, Palo Alto
118. Masters K (1985) *Spray drying handbook*, 4th edn. Wiley, New York
119. Burnett TA, Anderson KD (1981) Technical review of dry FGD systems and economic evaluation of spray dryer FGD systems. EPA-600/7-81-014. U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC
120. Lachapelle DG, Brown JS, Stern RD (1974) Overview of environmental protection agency's NO<sub>x</sub> control technology for stationary combustion sources. Presented at the 67th annual meeting American Institute Of Chemical Engineers
121. Srivastava RK, Hall RE, Khan S, Culligan K, Lani BW (2005) Nitrogen oxides emission control options for coal-fired electric utility boilers. *J Air Waste Manage Assoc* 55:1367–1388. Air and Waste Management Association, Pittsburgh
122. Hjalmarsson A-K (1990) NO<sub>x</sub> control technologies for coal combustion. IEACR/24. IEA Coal Research, London
123. U.S. Environmental Protection Agency (1992) Evaluation and costing of NO<sub>x</sub> controls for existing utility boilers in the NESCAUM region. EPA-453/R-92-010. Office of Air Quality Planning and Standards, Research Triangle Park

124. U.S. Environmental Protection Agency (1994) Alternative control technologies document: NOx emissions from utility boilers. EPA-453/R-94-023. Office of Air Quality Planning and Standards, Research Triangle Park
125. Heap MP, Folsom BA (1990) Optimization of burner/combustion chamber design to minimize NOx formation during pulverized coal combustion, Chapter 10. In: Pulverized coal combustion: pollutant formation and control, 1970–1980. U.S. Environmental Protection Agency, Air and Energy Engineering Research Laboratory, Research Triangle Park
126. Sommer TM, Jensen AD, Melick TA, Orban PC, Christensen MS (1993) Applying European low NOx burner technology to U.S. installations. Presented at the 1993 EPRI/EPA joint symposium on stationary combustion NOx control, 24–27 May 1993. EPRI/EPA, Miami Beach
127. LaRue AD (1989) The XCL burner – latest developments and operating experience. Presented at the 1989 EPRI/EPA joint symposium on stationary combustion NOx control, 6–9 Mar 1989. EPRI/EPA, San Francisco
128. Vatsky J, Sweeney TW (1991) Development of an ultra-low NOx pulverizer coal burner. Presented at the 1991 EPRI/EPA joint symposium on stationary combustion NOx control, 25–28 Mar 1991. EPRI/EPA, Washington, DC
129. U.S. Department of Energy (1996) Reducing emissions of nitrogen oxides via low-NOx burner technologies. Topical report number 5. U.S. Department of Energy, Pittsburgh
130. Campbell LM, Stone DK, Shareef GS (1991) Sourcebook: NOx control technology data. U.S. Environmental Protection Agency, Research Triangle Park
131. Wendt JOL, Sternling CV, Matovich MA (1973) Reduction of sulfur trioxide and nitrogen oxides by secondary fuel injection. In: Proceedings of the fourteenth symposium (international) on combustion. The Combustion Institute, Pittsburgh, pp 897–904
132. U.S. Environmental Protection Agency (1996) Control of NOx emissions by reburning. EPA/625/R-96/001. U.S. Environmental Protection Agency, Cincinnati
133. Chen SL, McCarthy JM, Clark WC, Heap MP, Seeker WR, Pershing DW (1986) Bench and pilot scale process evaluation of reburning for in-furnace NOx reduction. In: Proceedings of the twenty-first symposium (international) on combustion. The Combustion Institute, Pittsburgh, pp 1159–1169
134. Payne R, Moyeda DK (1994) Scale up and modeling of gas reburning. In: Moussa A, Presser C, Rini MJ, Weber R, Woodward G, Gupta AK (eds) Combustion modeling, scaling and air toxins. ASME FACT, vol 18. American Society of Mechanical Engineers, New York, pp 115–122
135. Lyon RK (1987) Thermal DeNOx, controlling nitrogen oxides emissions by a noncatalytic process. Environ Sci Technol 21(3):231–236. American Chemical Society, Washington, DC
136. Jødal M, Nielsen C, Hulgaard T, Dam-Johansen K (1990) Pilot-scale experiments with ammonia and urea as reductants in selective non-catalytic reduction of nitric oxide. In: Proceedings of the twenty-third symposium (international) on combustion. The Combustion Institute, Pittsburgh, pp 237–243
137. Muzio L, Quartucy G (1993) State-of-the-art assessment of SNCR technology. Topical Report No. TR-102414. Electric Power Research Institute, Palo Alto
138. Berg M, Bering H, Payne R (1993) NOx reduction by urea injection in a coal fired utility boiler. Presented at the 1993 EPRI/EPA joint symposium on stationary combustion NOx control, 24–27 May 1993. EPRI/EPA, Miami Beach
139. U.S. Department of Energy (2005) Selective catalytic reduction (SCR) technology for the control of nitrogen oxide emissions from coal-fired boilers. Topical Report Number 23. U.S. Department of Energy, Pittsburgh
140. Pereira CJ, Amiridis MD (1995) NOx control from stationary sources: overview of regulations, technology, and research frontiers, Chapter 1. In: Umit S, Agarwal SK, Marcelin Ozkan G (eds) Reduction of nitrogen oxide emissions. American Chemical Society, Washington, DC, pp 1–13
141. Nischt W, Woolridge B, Bigalbal J (1999) Recent SCR retrofit experience in coal-fired boilers. Presented at POWER-GEN international, 30 Nov–2 Dec 1999. New Orleans
142. Tonn DP, Uysal TA (1998) 220 MW SCR installation on new coal-fired project. Presented at the Institute of Clean Air Companies ICAC forum '98, 18–29 Mar 1998, Durham
143. Khan S, Shroff G, Tarpara J, Srivastava R (1997) SCR applications: addressing coal characteristics concerns. EPRI-DOE-EPA combined utility air pollutant control symposium – the mega symposium. Technical report TR-108683-V1. Electric Power Research Institute, Palo Alto
144. Foerter D, Jozewicz W (2001) Cost of selective catalytic reduction (SCR) application for nox control on coal-fired boilers. EPA/600/R-01/087. U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC

---

## PV Policies and Markets

WOLFGANG PALZ

Former EU Commission Official, World Council for Renewable Energy (WCRE), Brussels, Belgium

### Article Outline

Glossary

Definition of the Subject

Introduction

State of PV Solar Electricity Today: An Overview

Historical Political Development: Initiating a Market Explosion

Future Directions: PV's Promise in a 100% Renewable Energy Strategy

Bibliography

## Glossary

**Energy and power** Power is the capacity to generate electric energy. The power capacity is measured in Watts, kW (1,000 W), MW, GW, etc. The energy is measured in “Watt hours,” kWh, etc.

**Feed-in tariff, FiT or FIT** It is a tariff imposed by national legislation that grid operators have to pay for PV electricity delivered on their net by the PV generator operators. It is measured in cents/kWh. It is mostly employed in the European Union.

**Green certificates** In some countries such as the UK, some US states, and Belgium, the kWh of PV sold to the grid are paid with certificates whose value is defined by legislation. Those can be traded on the market place.

**Inverters** Inverters are electronic devices that convert the DC (direct current) generated in the PV modules into AC (alternating current).

**National REAP in the EU** In the frame of its Directive “Energy and Climate,” the European Union has agreed with all 27 EU Member Countries the REAP, i.e., the “Renewable Energy Action Plans.” They are mandatory obligations and must be implemented by 2020. In this frame each country has a particular target for PV implementation.

**Plus-energy buildings** Buildings whose energy consumption is more than compensated by internal energy production. This is generally achieved by means of PV integrated onto the buildings.

**PV cells, modules, systems** *PV cells* are made from silicon or other semiconductor materials. They are less than 0.5 mm thin. When interconnected by metal strips, cells are integrated into *PV modules*. They have typically a power rating of up to 200 W and may be 2 m<sup>2</sup> large. *PV systems* are power generators ready for use to produce electricity for a given purpose: feeding into an electric grid, supplying any appliance, in particular a whole building with power.

**PV: photovoltaics** Semiconductor plates that generate DC electricity when exposed to light, namely, the Sun’s radiation.

**Thin-film cells, a-silicon (a-Si), mc-silicon (mc-Si)** *Thin-film cells* are 100 times thinner than crystalline silicon cells. They do not come as individual cells but are integrated into larger modules during the

deposition of the thin semiconducting layers on a substrate. *Amorphous silicon* has no crystalline structure but one that is similar to glass; it contains a few percentages of hydrogen. *mc-Si layers* are made of silicon microcrystals in the range of a few nanometers. Layers of mc-Si are often deposited on top of an a-Si layer.

## Definition of the Subject

The large-scale generation of solar electricity by means of photovoltaics (PV) was traditionally considered excessively expensive and particularly problematic because of the Sun’s intermittency and the claim that the PV areas needed for deployment were prohibitively large.

Nevertheless, starting from a low level some 30 years ago, the global PV markets never stopped growing. In this very year 2011, the level of 60 GW of power will be exceeded globally for the first time. In the longer run, a major share of all the world’s electricity demand could eventually be supplied by PV. The reasons for the new market success of PV are manifold and are presented in detail in this article.

## Introduction

Large-scale power generation by means of photovoltaic conversion of the Sun’s rays into electricity is a newcomer in the global electricity markets. The underlying reasons for this new interest are associated with the growing concern about the global effects of climate change and additionally with the challenge of security of energy supply. This has recently been illustrated again by the problems and new concerns raised after the nuclear accident in Japan and the threat of continuously growing oil prices on the world markets.

PV is one of the solutions to address these problems. The Sun that supplies the fuel for PV power generation is accessible everywhere. In concert with the other forms of renewable energies, PV can play an ever-growing role with the prospect of achieving ultimately a 100% renewable energy supply worldwide.

The intermittency problems of PV power can be solved in combination with hybrid power supply systems associated with the development of electricity storage in batteries and other devices resulting from the increased interest in electric cars and electric transport in general.

Further, PV has the particular interest of making it possible to go progressively to a more decentralized electricity supply system and reducing the needs for larger transport and distribution networks of power.

The entry starts by providing a comprehensive update on the PV's status today around the world. Emphasis is being put on the tremendous developments that occurred in particular during the last 2 years, 2009 and 2010. Investments in new module production lines in the GW range have induced tremendous decreases of costs and prices and consequently an explosion of the world markets.

A particular analysis is addressed to the national market support schemes. The reasons why "PV tariff" systems – involving many billions of \$ – are positively viewed in some countries and in others not are for the first time analyzed in detail in such a publication.

Before attempting an outlook in the short and longer term future, the entry goes back for a moment to recall the historical events and the political fighting that was needed to kick start the market growth that the world has seen since the early years of 2000.

The article is closed by a comprehensive overview of PV references with emphasis of the many currently available sources from publications on the Internet.

## State of PV Solar Electricity Today: An Overview

### Global PV Markets

In 2010 a PV power capacity of 18,000,000 kW or 18,000 MW, i.e., 18 GW was newly installed. This is reported by the British Market research company IMS [1]. For the first time ever, more than 10 GW of PV capacity were globally installed and connected to an electric grid in one single year. 18 GW represent an overall collector area of over 120 km<sup>2</sup> – as 1 kW of "peak power" corresponds typically to a module area of some 7 m<sup>2</sup>.

They stand for a global market volume of 70 billion USD or 50 billion €, just for the modules. On average the PV modules make up for 50% of the total cost of a turnkey system. When including the balance of system, installation costs, etc., global PV business attained a total market value of 100 billion € in 2010.

Over 80% of the global new PV capacity was installed in Europe. For the first time Europe set up

more PV than wind power that year and more than any other form of conventional power.

To measure what has been achieved in the last 30 years it should be reminded that in the year 1980 world sales of PV amounted to just 4 MW [2], that is 4,500 times less than now. During the last 30 years the global PV market never stopped growing but with a rate of 130%, last year was also exceptional from this point.

IMS [1] announced also for 2010 a global PV cell and module manufacturing capacity of 34.5 GW in place. But only a part of it was in full production. A total area of 22 GW were actually produced, of which next to the 18 GW that were sold in the markets, some 4 GW went to inventory.

With *annual capacities* of 7.4 GW, 3.5 GW, and 1.36 GW installed during 2010 Germany, Italy, and the Czech Republic were globally number 1, 2, and 3 on the global hit list. All other nations including the USA had less than 1 GW newly connected during 2010. In summary this is the following ranking of new PV installations for the single year 2010:

1. Germany 7.4 GW
2. Italy 3.5 GW
3. Czech Republic 1.36 GW
4. Japan 991 MW
5. USA 878 MW
6. France 720 MW
7. Belgium 355 MW

This is a snapshot for 2010: It is already clear by now that this order is going to change in the future because of the restrictions in some countries that are similar to those Spain had taken in 2009.

The *cumulated* PV capacity installed worldwide looks like this:

In total a world PV capacity of over 38 GW is by now in early 2011 connected to the electricity networks. With this, some 6 GW of stand-alone PV must be included in the global account [3]. In total there was at year's end 2010 some 44 GW of PV in all sizes installed worldwide. As it is expected that more than 20 GW will be newly added worldwide in 2011, the global PV capacity will have reached 65 GW in early 2012.

Almost half of all globally grid-connected PV was in place in Germany: 17.3 GW. Next to Germany, Italy

stands out as a brilliant second with a total of 7 GW installed – even though part of it was not yet grid-connected at the end of 2010 (Although the orders of magnitude were different it is interesting to note that in the 1990s the national R&D budgets for European PV came in the same order: Germany first and Italy second [4]). The Czech Republic had a *cumulated* capacity of 1.82 GW at the end of 2010.

On March 22, 2011, Germany had at one time more PV capacity feeding electricity to the grids, i.e., 12.1 GW than all nine atomic power plants in operation with 12 GW; it was a breakthrough.

The market in Germany favors decentralized PV deployment: In total 860,000 individual PV generators were in operation there at the end of 2010. Of them 230,000 have been installed just in the last year [5]. Fifty eight percent of the newly installed PV capacity went for systems with a capacity under 100 kW – most of those had a capacity between 10 and 100 kW [6].

In Italy 250,000 plants were widely distributed over the country at the end of 2010 [7].

Of Japan's 991 MW PV market in 2010, an overwhelming majority was installed on 200,000 residential households (803 MW) [8].

In early 2011, there is a total worldwide installation of well over 1.5 million individual PV plants connected to the electricity grids.

Even though a high number of PV plants were attached to a building envelope, the proper "building integrated PV, BIPV" in the architectural sense remained in an early stage of development. Keeping in mind that in particular in Europe national policies are moving now in favor of the so-called plus energy buildings and hundreds of thousands of them are going to be built, BIPV will have a great future: A building's energy balance where it produces more energy than it consumes is difficult to achieve without proper integration of PV.

It is true to say, however, that currently the ground-mounted utility-scale projects take a considerable share of the global PV market: In the USA about half of the capacity installed in 2009 was for system sizes of over 100 kW [9]. The reasons for this preference for larger plants are in general, among others, local bureaucracy and the kind of administrative regulations that make the procedures for small systems cumbersome. The same was true when Spain had its big PV boom before 2009.

The largest single PV project identified in 2010 [10] was Sarnia with 97 MW in Ontario, Canada; not far behind came Montalto di Castro in Italy with 84 MW. In Arizona, a 290 MW PV installation is currently being built in Agua Caliente, and in China are talks in progress of a 2 GW PV system that would be built in steps in the next few years.

PV power plants are indeed unique as they cover a range of over 12 orders of magnitude of capacity: mW for watches and pocket calculators, some Watts for street lights, solar-home systems, SHS, in the poor villages, kW for building integration, and MW and GW, the sizes of conventional centralized power plants.

China currently dominates the world market of PV modules: half of the PV modules sold in 2010 were manufactured in China; plus some 8% in Taiwan [11]. This impressive proportion did increase all the time over the last few years. Virtually all of it is exported as China has by now no significant home market for PV.

What are the reasons for China's competitive edge in current world markets? This author believes that it is certainly true that the Chinese industry is very effective in bringing products of high quality to the market; but the essential reason for its market dominance seems rather to be the undervaluation of the Chinese currency, the Yuan. In the end this very situation was – together with the national support schemes for PV in some countries – the driving force that triggered a revolutionary decrease of cost and prices on the world's PV markets and correspondingly the developments of these markets as have been seen over the last few years.

As could be expected, many nations such as the USA and France are unhappy with the dominance of imported PV products on their home markets. Some try to react with "local content rules" imposing that 50% or more should come from domestic production [12]. Germany being a world's leading export nation, makes a point of following a liberal route and accepts all the PV imports: But still, German PV industry has a strength of its own and was able to keep a share of 40% in its domestic market while 50% came from China [13].

In terms of technology, thin film modules had a share of 13% of global PV market in 2010 while a year earlier that share stood at 17% [11]. This may

look disappointing for the “thin-film enthusiasts” but it still means an increase in absolute terms. The reason for the ongoing dominance of crystalline silicon in the markets has to do with the vitality of the Chinese producers who are not so much interested in thin films for the time being; another reason is maybe that it took more time than previously thought to bring cells of the young CIS/CIGS family from laboratory to mass production.

### **PV's Role in General Electricity Supply**

A PV array of 1 kW will typically generate an electric energy of between 1,000 and 2,000 kWh over a whole year; in the winter months production will generally be about half of what can be expected in the summer. Details will obviously depend on the climate at the place of installation, if the array is fixed-plate or tracking the sun, and other constraints. As an order of magnitude, a PV generator produces per kW and year in terms of energy only half of that of a wind turbine at a “typical” wind regime. It also produces less per kW than most conventional power plants.

But what counts in the end is the kWh cost.

In that respect, latest reports in the media tend to become most optimistic about PV's competitiveness with some of the key electricity providers of today: two scientists of the Duke University in North Carolina came recently to the conclusion that PV is indeed getting competitive with modern nuclear plants [14]; also in the United Kingdom an industry participant has claimed that the proposed atomic plants foreseen for construction in the country will not be able to match the price of solar electricity [15]. Concerning the outlook for competition with natural gas – arguably the cheapest electricity in the market today – it is noteworthy that the utility Southern California Edison filed recently the request of approval for a total of 250 MW of PV at lower price than electricity from combined-cycle natural gas plants [16].

With respect to other renewable sources of electricity, too, “expensive PV” as conservative politicians like to call it, started recently to look a lot more attractive: The cost information about “concentrating solar plants CSP” is traditionally much more opaque than that for PV, but the message is eventually spreading that CSP's promise should be seen with caution [17].

A significant revolution with respect to wind energy occurred also in 2010 in Germany: 7.4 GW of PV were installed in the country and only 1.55 GW of wind power. Traditionally the relation used to be the other way around. Some years ago when Germany was a world leader for wind power, PV deployment was yet at an embryonic stage. As far as off-shore wind is concerned, latest cost figures also indicate that PV is getting competitive here faster than previously thought.

As of today a relevant region with a particularly high contribution of PV to its overall electricity consumption is the German “Free State of Bavaria”: With over 6 GW of installed PV capacity Bavaria has since 2011 exceeded a 7% PV electricity supply to its overall consumption. By the end of 2011 this proportion will no doubt go to 10%: When buying a BMW or an AUDI car that are actually Bavarian products, the fact is that solar energy has already now a nice share embedded in these cars.

Germany as a whole draws now on average 3% of its electricity demand from its PV plants. On favorable days the PV plants produce more than 8 of the country's 14 nuclear power plants [18]. Other countries with significant PV contributions of over 2.5% to their national electrical consumption are Italy and Spain; all other nations stay for the moment well below 1%.

As to transport and distribution of electricity, PV plants are currently well integrated into the low-voltage grids: thanks to local production and consumption, transport over long distances must not be necessary. In that respect PV plants are different from most conventional ones. Even for wind power, long-distance electricity transport is often essential as the consumption centers do not always coincide with the regions of sufficiently high-wind regimes, namely, when offshore wind on the sea is being talked of. For PV that is much less an issue, even though in Germany in areas of high PV deployment the capacity of the medium and low voltage distribution grids may currently become stressed and may need reinforcement. . .

### **National Policies and the Role of Financial Support**

In 1983, it was written [2]: there is a “chicken-and-egg situation: new technology such as PV will not reach large-scale, low-cost commercialisation unless a market

is perceived by the producer. . . . Manufacturers learn to make a cheaper product as output increases. . . . The problem is particularly acute for high capital cost/low running-cost systems such as PV. Another paradox is that, even when economic break-even cost levels have been reached potential customers may still decide to wait in anticipation of further cost reductions. . . .” It was also written: “The only alternative would be for governments to provide the necessary support and turn what is technically feasible into an established fact.”

Another anomaly with respect to PV deployment of those days was the trend to offer “clean solar electricity” at real cost. This cost was and still is obviously higher than the market price of conventional electricity from fossil and atomic sources, but as people declared they were ready to pay more if the electricity was clean – it looked for some promoters as an option. Even though there was never a level-playing field in the electricity market. Subsidies are distorting all market prices in this sector, above all those for the conventional energies: the International Energy Agency IEA in Paris declared that in the year 2009 global subsidies for the fossil energies had been 312 billion USD – not mentioning those for nuclear electricity – compared to 57 billion USD for all forms of renewable energy of which PV was only a small part in that year. In particular from an ethical point of view, higher market prices for solar electricity are an absurdity; it cannot be that one pays less for polluting and unsustainable electricity than for PV: PV has a wealth of environmental and social merits together with those of unlimited supply that make it institutionally desirable everywhere. The external costs of conventional electricity production and supply do have to be included in market prices to compensate what economists call a “market failure” [19].

The market explosion for PV started actually in 2004 with the introduction of the “feed-in tariff, the FiT” in Germany. Elsewhere in this paper the details of the political decision process in those days would be discussed. No doubt, the FiT is at the origin of a global PV market of some 60 GW connected to the grids at the end of 2011. Still in the 1980s and 1990s nobody would have expected such a thrilling success. And in all logic it is currently Germany where the political initiative started that leads the world markets.

The FiT in *Germany* is actually not a subsidy: this was a decision by the highest European Court in

a dispute with the German electric utilities. The reason is that the favorable tariffs for the PV electricity that is supplied to the grids are paid by the grid operators and are not coming from public financing sources.

The situation may be different in other countries such as for instance Spain where electricity is anyway partially subsidized by the government and where the tariff payment for PV comes on top of the ongoing subsidy.

The essential features of the FiT are the following: (a) the grid operators have the obligation – unless it is technically just not feasible in a certain situation – to buy any kWh from PV (and other forms of renewable electricity that are specified in the law) that is offered to them at any time; (b) the operators have to pay from their own budget the particular tariff that was decided previously by the government.

There is a large consensus across German society and the industry that the FiT legislation or “EAG EE” as it is called after the latest revision in March 2011 has great value: It allows for reasonable planning and administrative approval schemes, a stable and accountable policy, and last not least attractive “returns on investments” of 8% and more. The FiT system has only winners; everybody gains money with it: the investors, the banks, the communities, the PV industry and its commerce. The money inflow comes from the electric utilities, i.e., eventually from all electricity consumers – with the exception of some industries. In early 2011 all customers, with their exception, contributed for some 8% of the tariff applicable to all electricity they bought from the net. But in reality the part of the FiT in everybody’s electricity bill is only a smaller fraction of all the extras the utilities have added to these bills: between 2002 and 2011 the tariff for consumers was increased by 12 € cents of which the FiT part represented just a quarter [20].

The share of the FiT in the general tariffs is expected to stabilize at about 8% in the future: While the number of plants in the country will keep growing, the FiT is decreasing by decision of the government in line with the decrease of the PV module and system prices (the PV FiT is being cut by 50% between 2009 and 2012) – that is the new agreement reached between the German Government and the PV industry. This decision is a breakthrough as in the past PV investors in Germany made excessive profits when the module prices

decreased faster than the feed-in tariff; that happened in particular in 2009. Since the beginning of 2011, the FiT regulation will be frequently adapted just in phase with the development of the German PV market volume. In this way it can be avoided that profits together with the market volume grow excessively: the target is to keep the market below 5 or 4 GW/year.

As mentioned before, one of the beneficiaries of the PV deployment in Germany are the municipalities: in 2010 the value created in villages, towns, and cities by PV through tax and leasing revenues, corporate earnings, jobs, and saving of conventional fuels reached a new record of 5.8 billion €.

At the beginning of 2011 Germany had 133,000 jobs associated with production and deployment of PV. Half of them were linked to various suppliers and module manufacturing; the rest was for installation services, mechanical engineering, etc.

PV plays also a considerable role in Germany's regional development. South of Berlin, in the declined economies of what was previously Eastern Germany's GDR, a "Solar Valley" has developed as a brilliant success of the country's effort to make the regions' economies "blossoming" again. Most German PV module and silicon feedstock manufacturers have settled in the "Solar Valley." Here a win-win situation is seen: Whereas the regions in difficulty get enormous new hope to see their economies developing again, the PV industry benefits from low labor cost that prevails in these areas. In addition, the industry benefits from considerable subsidies for their manufacturing plant investments on behalf of the European Union's "Structural funds."

Since the revision of the German FiT regulation in 2009, the ecologists are a lot more happy as new plants can no more be built on agricultural land.

Another interesting addition was the purpose of promoting self-consumption by a storage system: When more than 30% of the PV electricity produced is consumed "in-house," the paid tariff is increased by 8%. It is a first step, but a lot more would be necessary to stimulate the combination of storage capacity with the PV installation to weaken the role of the electric grid and its domination of the whole deployment structure.

A remaining deficiency of the German FiT is the fact that the PV tariff is presently not modulated

following the regional differences of solar irradiation in the country. Most PV plants are now installed in the South where the income is higher. Typically, Bavaria in the South had in 2010 four times the PV capacity installed in Lower Saxony in the North.

To complete the facts and figures on the German market leader for PV one should mention two additional financial benefits.

The investment of PV plants is supported by favorable credits from the state-owned KfW bank. Since February 2011 the interest rate of a KfW credit for PV plants stands at 3.5% with a duration of 20 years, of which 10 years at a fixed rate and 3 years amortization free [21].

Building integrated PV systems can be amortized by tax credits over 20 years [22].

Interesting innovations have recently been offered by the German industry for combining PV with heating of the building. Some propose a combination of PV, solar heat collectors, and intelligent windows. Others propose systems that make use of PV rejected heat for house heating by means of a heat pump.

In summary, there are many reasons why Germany has become a market leader of global PV. First of all, PV has gained support, more than any other form of energy – even among the renewables – across all parts of German society. Support remained unchanged when the Federal Government changed in the mid of the last decade from social democratic "left/green" to conservative "right/liberal." There is no prevailing complaint about an excessive financial burden for PV like in other countries.

The government made it clear that PV is a clear winner for everybody and society at large when including the external effects of energy generation and supply. For 2009 the net cost of PV support amounted officially to less than 6 billion €. But the financial benefits to compensate it were actually a multiple of that cost. They are listed as follows:

- Billions of euros saved on environmental degradation
- Carbon certificates gained in the European carbon market
- 5 billion € saved on energy imports
- 4 billion € for "Merit order," a cost saving on the European electricity spot market in Leipzig. It takes



account of the effect that at certain times renewable electricity is available at the lowest cost of all electricity on offer; the price difference with the next dearer “conventional electricity” is the “Merit Order”

- 5 billion € as value created in the municipalities through tax income and others
- Value created through export of products and services, 130,000 new jobs, technological stimulation

It is said that the FiT was adopted by over 60 nations by now. But in most cases countries were facing considerable teething problems to get their legal regulations right. In their new enthusiasm for PV, many legislators were too generous, put the PV tariffs too high, added very favorable tax regulations and found themselves eventually confronted with mountain-high financial burdens of billions of euros to be borne by the national budgets. No wonder that investors had been rushing into the new business where return on investments could exceed 20%. It happened first in Spain in 2008/2009; France, Italy, Czech Republic, the UK, and others followed. Emergency measures had to be taken by the legislators. The options were either to impose a cap on the yearly PV capacity to be financially supported for installation and grid connection or to decrease radically the PV tariffs and the tax credits, or to do both.

And to make things worse, investors had to go through a national tendering process to get approval.

Spain and Czech Republic even imposed retroactive restrictions.

*Italy* had the second biggest PV market worldwide in 2010. There had been discussions, however, if really 3.5 GW had been installed in that singular year to reach a total of 7 GW. There were complaints about fraud, a possible involvement of the mafia that had already intervened previously in the Italian wind market; the question was raised, too, if not a major part of the PV installations were not yet grid connected. After discussion between the government and the Italian PV associations, new measures will no doubt make the national support system more sustainable. Measures in the pipeline are an annual cap, a reduction of the PV FiT down to 25 € cents/kWh, auctions, etc., [23].

*Spain* has almost 4 GW of PV capacity installed, but since 2009 the national market came almost to

a complete stop. A 500 MW annual cap and a new auction mechanism for future contracts were decided. The PV FiT is being strongly reduced. As a result, PV industry in Spain is losing companies and up to 40,000 jobs. The background for the restrictions is the fact that Spain has since 2002 set up 22 GW of combined-cycle gas plants. As the demand of electricity has weakened, these plants are partly lying idle. To ease the overcapacity, Spain is exporting electricity as much as it can. At the end of 2010, the government has taken emergency measures: While freezing electricity prices for the five million poorest families, the incomes of all PV plants in the country are trimmed by 30% – for a duration of 3 years. The reduction is achieved not by reducing the FiT height but rather by paying it for a lower number of hours than the kWh produced over the year.

*France* had 720 MW of PV newly installed in 2010 – almost three times more than California that year. 152,000 PV systems were set up. France used to be a pioneer for rooftop PV installations; they had a considerable bonus in the financial support system.

By the end of 2010, construction and grid connection permits of 3.4 GW had been issued; then the government imposed a moratorium. As a result most of these systems were not going to be built. In March 2011 dramatic changes in the French PV regulations were introduced: an annual cap of 500 MW was imposed; systems of 100 kW and more are subject of a tendering procedure and they will receive not more than 12 € cents/kWh; the FiT for smaller systems were reduced by 20% immediately in March and a further 10% every following quarter to reach a reduction by 40% end 2011.

Talking about reductions and restrictions one must not omit the decrease of the French tax credit for private investors in PV that went from 50% to 22.5%, effective 1 January 2011 [24].

*The United Kingdom* is a latecomer for PV promotion by FiTs. The market still is very small. But Britain did not want to stand back in the recent wave of PV market restrictions. PV systems over 50 kW will see their tariff capped from 30 pence/kWh to between 8.5 pence and 19 pence/kWh [25]. . . .

*Japan* has a very long record in PV promotion but it is another latecomer in terms of PV promotion by means of the FiT scheme. The Japanese FiT is called “Excess Electricity Purchasing Scheme”; it stood in

early 2011 at 42 Yen/kWh. To better promote MW-scale PV plants, the PV FiT for such systems was increased from 24 to 40 Yen/kWh [8]. PV is also locally promoted in Japan by the “School New Deal” project by the Ministry of Education, by introduction measures by METI, by local governments, etc.

A special case among the European countries in terms of PV market support is “little” *Belgium* in the heart of Europe. It has a market promotion system of its own that is not linked to the FiT. It is similar to those in the USA, but for the time being more successful. Hence, it is worth a closer look.

Belgium has installed in 2010 355 MW, i.e., 40% more than California that year [26]. 95% of all installations were residential; Brussels, the capital of Europe, had 2,000 roof-mounted PV systems grid connected. Many industrial promoters are in competition on the Belgian PV market. They guarantee a rate of return on investment of 8%. The profitability comes from the conventional electricity that is saved, a 40% tax credit on the investment, and the income from “green certificates.” The regional governments pay a number of certificates in units of 1,000 kWh fed into the grid. Those have a minimum value that is increased by a trading system in which the electric utilities buy those certificates as needed for achieving the quota of green electricity imposed on them.

In late 2010 the Flanders Government decided to decrease the value of the certificates they issue in 3-month steps by some 10% each time.

Self-consumption gets promotion as systems over 1 MW with less than 50% self-consumption will receive green certificates of much lower value.

It is also interesting to notice that the Belgian system includes a bonus for low-income investors: for building integrated PV the subsidy for such families can reach 1 €/W up to a total of 30% of the installed system cost.

The USA has a long record of PV development and enthusiasm for its deployment. However, by international standards, the current American PV markets as of 2011 are relatively small. 878 MW had been installed in 2010 at a market value of \$6 billion [27]. In the USA the FiT scheme is not employed, but as shown for the example of Belgium where it is not employed either, the FiT alone cannot explain the difference.

The American PV market environment is obviously more complex than, for instance, the German one.

PV system prices on the markets are blown up through numerous effects.

It is mentioned that in the US “balance of system” costs are twice of what they are in Germany; even module prices are higher there so that American installers are not fully benefiting from the favorable price developments on the global markets. “Permitting issues and delays resulting from the USA’s jurisdictions” are problems that cost money [28]. Access to critical financing is also a difficulty.

American support for PV includes several original elements. Firstly, there is the “net metering” that allows delivering PV electricity into the net at the same price as the usual sales tariffs to customers. Secondly, there was still in 2011 a federal 30% cash grant for all investors in communities, from banks, private parties. Thirdly, the renewable Portfolio Standards RPS in the States. It is a requirement on retail electric suppliers to supply a percentage of their retail load with PV or another form of renewable electricity; they are often associated with a tradable renewable energy credit, REC. Each state has a different arrangement for it.

And much of the remaining support is linked to tradable tax credits in various forms. “Many developers are borrowing against the cash grant to raise money for construction, then handing the payment over to the construction lender when the project is completed” [29]. “The off-take agreements, the ‘power purchase agreements’ or PPA are essential to lock in revenue streams from the power plants, and lenders will not consider projects that do not have a PPA in place” [30].

“Bloomberg New Energy Finance” estimates that commercial-scale PV systems can obtain returns of 8–14% in certain states.

One may also mention the subsidies given by the Department of Agriculture to farmers and ranchers to install PV and other clean energies.

In California there are new initiatives to install PV for the benefit of low-income families.

Last not least, a note about *China*, the giant in PV module production with 120,000 PV jobs in place [31], and almost a dwarf in domestic installations: 0.8 GW existed in total at the end of 2010. So far China was lacking a PV promotion scheme that worked.

In March 2009 the government launched the “Golden Sun” Programme. It has as far as one can know two routes.

One concerns the setting up of larger system blocks in the west of the country. After tendering, just a few hundred MW are supposed to be built for an incredibly low prices down to 8 € cents/kWh [32]. One will have to wait for implementation to see if this is realistic.

The second approach looks very original and promising; it focuses on PV implementation in the “nonresidential sector” [33]. The idea is the following:

- In industrial and commercial zones, administrative quarters and educational buildings energy demand is better in phase with solar generation during the day; no net-metering and feeding into the public grid is necessary, all PV generated electricity is for self-consumption.
- In China currently the kWh price from the grid is twice as high for the industry than it is for residential use. For Europeans this is surprising as in Germany, etc., the national industry benefits from the lower electricity tariffs for protection and many of the companies are not even sharing the cost of the FiT like all the residential grid customers do. Hence it is a fact that in China the incentive for PV utilization is higher in the nonresidential area
- The government provides the PV modules and the inverters at half price with a 50% subsidy
- The government buys the modules in a large tendering exercise. Central procurement will further lower the PV module prices
- All PV systems are to be quality controlled and certified after installation

Also impressive are the favorable credits of some 20 billion \$ that the Chinese banks provide to the Chinese module manufacturers.

### PV Industry, Products, Costs, Prices

The following module manufacturers were leading the global markets in 2010 [34]:

1. Suntech Power (China) 1.57 GW
2. JA Solar (China) 1.46 GW
3. First Solar (USA) 1.41 GW
4. Q-Cells (Germany) 1.01 GW
5. Motech (Taiwan) 0.85 GW
6. Gintech (Taiwan) 0.72 GW
7. Sharp (Japan)
8. Kyocera (Japan)

They all sell crystalline silicon cells and modules, except First Solar which is the market leader for CdTe thin-film solar cells, and Sharp that markets next to crystalline silicon also amorphous (a-Si) and microcrystalline silicon (mc-Si).

The ranking is actually rather volatile. Years ago, Sharp was for a long time the global leader; later it became Q-Cells. The latter went through a crisis when in one single year it accumulated a deficit of 1.5 billion €. First Solar took the helm in the following year 2009.

In terms of technology, a provisional ranking came for 2010 from Paula Mints of Navigant Consulting [35]:

- Monocrystalline Si +/-43%
- Multicrystalline Si +/-43%
- CdTe 9%
- CIS/CIGS 2%
- a-Si/mc-Si 2%

There is consensus among experts that the family of CIS cells, the “Copper Indium Diselenide,” is particularly promising among the thin-film cells. Numerous companies worldwide are working on it. The latest example was the announcement of Solar Frontier, a company of the Shell Oil Group in Japan, in last April 2011 that it opened a new factory for CIS modules in Japan. It targets a manufacturing capacity of 1 GW within months.

In general terms, there is no type of cell or module that could be given preference for matters of principle.

GaAs cells have the highest efficiencies: Lately over 43% have been achieved. But because of their high cost they only found a market on space satellites and on concentrating PV systems, the CPV.

Crystalline silicon cells, as far as they are concerned, offer today still higher conversion efficiencies than thin-film cells; over 18% are now readily available for Si modules on the commercial markets. The American SunPower is the leader with up to 25% efficiency. It is not unfair to say that the efficiency development on silicon cells did not see a major breakthrough in the last 40 years: in 1972 the USA already had the “violet cell” with 16% efficiency and 2 years later the “black cell” with 19% [36]. On the more negative side, one could mention the higher complexity of the silicon technology and hence a somewhat higher production cost.

Most thin-film cells are a factor of 100 thinner than the massive silicon ones implying a gain in material

consumption and cost. Today's market leader is CdTe. Typical efficiencies for this material go up to "only" 12%. It is interesting to note that a very thin additional intermediate layer of CdS does contribute to this efficiency [37]. The simpler thin-film technology leads to attractive manufacturing costs: First Solar announced one half € per Watt. But this cell has its own enemies because of the cadmium it uses. In particular in Japan, this raises a problem as the country had a bad record with cadmium pollution in the past.

No wonder then that – as mentioned before – Japan has turned its attention toward the CIS family of cells and a-Si cells. Among the many labs currently working on CIS, the Japanese company "Solar Frontier" mentioned before has claimed a record efficiency of over 15%. NREL in the USA has recently measured 15.5% on a  $30 \times 30$  cm CIS cell of the US company Avancis. The problem looming around the corner is the "scarcity" of indium in nature; it is an essential compound in the cell. It could become a problem in case of mass production.

Finally there is a-Si and mc-Si. A new record has been reported here also by NREL, the US Institute which is an international reference in this field. On  $400 \text{ cm}^2$ , 12% efficiency has been observed. Since the days of tremendous stability problems linked to the Stabler–Wronski effect in hydrogenated a-Si, there has been great progress thanks to the combination with mc-Si layers, a newcomer developed in Switzerland. This type of cells is now also among the promising ones.

Many cell and module manufacturers today are "GW" companies. As the economy of scale is the determining factor for achieving competitive and low-cost market products, smaller factories are disadvantaged. All experts, from the early days on, were unanimous to see the future for PV in large GW-size production [38].

From this short list of global module manufacturers, it might be wrong to hastily conclude that just these few companies dominate the entire PV business. Typically, a company such as SolarWorld is not even included on the larger list of 12 module manufacturers even though it is highly profitable and a darling on the stock markets: It is one of those "vertically integrated" PV companies that put under one roof all business from silicon feedstock to plant installation.

There is also a large fraction of PV companies specializing in silicon material production, such as Hemlock in the USA, Wacker in Germany, and many

others in Norway, China, Japan, Korea, etc. The global silicon feedstock capacity was estimated at 0.23 million tons in 2010 [8]. Nowadays more silicon goes into PV module production than in the electronic chip business. The business exceeds globally the 15 billion € mark.

A major part of the global PV business involves the specialized equipment providers, those who sell turn-key automated cell and module manufacturing plants. An example is the German Roth & Rau that just recently changed ownership. Solarbuzz [39] mentions for 2011 a global business of over \$15 billion for the equipment needed for a capacity expansion of the ingot, wafer, cell, and module manufacturing. Thirty percent is expected to go for crystalline silicon and the majority of 70% for thin-film modules. Almost 80% of this year's thin-film investments might go to the CIS family of technology and the amorphous/micro-crystalline silicon group of materials and structures.

At the end of 2011 a new production capacity record for a single company is expected: 3 GW for silicon cell producer JA in China.

And last not least, there is a global \$6 billion market for solar inverters. Very high DC/AC conversion efficiencies above 98% are international commercial standard today. One of the leaders is SMA in Germany that has a long-standing record for specializing particularly in this sector.

To conclude this chapter, *PV's market economics* should be briefly reviewed. In that respect it is convenient to look at the case of Germany. Its solar resource is relatively low: What holds true for Germany then is also applicable for most other regions that benefit from a more favorable solar irradiation. On the other hand, the FiT tariff is a good image of the PV's market value. If it were not profitable at a certain FiT value in place, investors would certainly not go for it.

At the end of 2010 one could find on the Internet in Germany the following wholesale prices per Watt for solar modules, all TÜV certified:

- CdTe modules 1.38€, 80 cents on the spot market
- Si modules from China 1.55€, 1.24€ on the spot market
- Si modules European origin 1.75€

As mentioned before, it is no secret that the cost advantage of the Chinese modules reflects very well the

undervaluation of the Chinese currency with respect to the euro.

System prices (without storage) varied between 2.10€ and 2.90€; they were the lower, the larger the system. On the spot market prices down to 1.60€ could be found. In 2006, those prices stood still at 5€ in Germany.

Since early 2011, the German FiT was set at 28.74 cents/kWh for building integrated systems and 21.11 cents/kWh for ground-mounted ones.

Two conclusions follow:

1. The PV electricity cost from large in-field systems is currently still two to three times higher than electricity from the high-voltage grid. It will need many more years to get competitive in pure monetary terms. This will be achieved for a smaller part because of lower PV costs and for a larger part as a result of the need for more “sustainable” electricity on the networks and hence higher commercial costs there.
2. PV looks a lot more favorable for decentralized power supply. Consider this: The German FiT is expected to be lowered in the course of 2011 to  $\pm 25$  € cents/kWh while the average cost of conventional electricity sold to every consumer will this year increase from 23.70 cents/kWh by 8%. In conclusion, the crossover of both – the German “grid parity” – will be at around 25 € cents and it is going to happen in the course of this year 2011.

### Historical Political Development: Initiating a Market Explosion

To stimulate the PV markets, political authorities have first tried the so-called Solar Roof Programmes. The German Government put the first one in place in 1989/1990; the responsible official in Bonn was W. Sandtner. In 1994 the Japanese Government’s Cabinet decided another one that was larger in size ( $\pm 70$  MW). A third one, a 100,000 Roof Programme was eventually decided in Germany by the Minister of Finance O. Lafontaine, the KfW Bank, and a few Parliamentarians, namely, H. Scheer, H-J Fell, and others. It started on 1 January 1999.

The driving force in Germany was the late Hermann Scheer, Member of the Bundestag. All that followed

was, as H. Scheer called it, a “High Noon Story” until the real thing, the FiT was in place in Germany in 2004, with over 60 other nations following suite since then – and the market explosion that has been seen.

The whole story is explained in Ref [38], its Preface of H. Scheer and on page 77. It will only be summarized in the following.

The German PV Roof Programme was a zero-interest scheme in which these interests and also the final installment were borne by public finance. Contrary to the later FiT, investors could not count on any profit, on the contrary. It was just a subsidy for less than one third of the total investment.

As a next step, the Bundestag decided – after a harsh fighting – an additional “feed-in-tariff” of 99 German Pfennige (about 50 € cents) – but only for the investors of the PV Roof Programme. This Act became effective on 1st April 2000. On this, the EU Commission in Brussels filed a law suite against Germany at the European Court of Justice. In March 2001 the law suite was dismissed.

This arrangement of a FiT paid on top of the investment subsidy lasted until 2003 when a total of 300 MW had been newly installed in Germany. As the 100,000 Roof limit had actually been exhausted at that point, it was time to establish a simple FiT on its own right, called EEG. Before this was going to happen in 2004, the preparatory process was bridged by a “Transitional Law” put forward by Hermann Scheer and eventually imposed by the Parliament against the will of the responsible Minister J. Trittin.

Eventually, the EEG including the FiT was decided in April 2004. That was the kickoff date for modern PV.

The value of some PV companies on the stock markets soared from one day to the other. Since then many people around the globe – not only in Germany – became PV millionaires.

To conclude this story it is important to mention that a cap of 750 MW always existed in the German FiT; it was eventually removed only in 2009, i.e., just 2 years ago.

It is interesting to add at this point a note on certification of PV products. For the FiT and the US or Belgian “green certificates” scheme, certification is being taken care of by the inherent interest of the investors: If the investor does not make sure he got a certified product, he will lose an income with the

kWh payment. Not so with the PV subsidies for roof-mounted systems. Here an additional certification must be superimposed. If not the subsidy sponsor may be the loser.

### Future Directions: PV's Promise in a 100% Renewable Energy Strategy

The European Union, with a population of 500 million in 27 Member Countries, is globally the only major region or nation that has a precise and binding target for the implementation of the renewable energies by 2020. The legal instrument is the EU Directive "Energy and Climate."

The EU has the plan, on average over all its Member Countries, to double the renewable energies' share from 10% in 2010 to 20% in 2020. The EU Member Countries have defined their own particular shares inside this target in the so-called national Renewable Energy Action Plans or "nREAPs." Details can be found on the EU Commission's web site or [40, 41], and others.

Following the nREAPs, the EU's share of electricity consumption from renewable sources like hydro, solar, wind, etc. will increase from 19% in 2010 to 34% in 2020.

Within the frame of the nREAPs, the EU projects a cumulated PV capacity of 84 GW within its borders by 2020, more than half of it in Germany.

The hit list of the eight leading countries is the following:

- Germany 51.7 GW
- Spain 8.4 GW
- Italy 8.0 GW
- France 5.4 GW
- UK 2.7 GW
- Greece 2.2 GW
- Czech Rep. 1.7 GW
- Belgium 1.4 GW

These figures were put together in the course of 2010 by the bureaucracies in the national governments. In some cases, the targets are already obsolete by now. In particular after the nuclear catastrophe in Japan, there are reasons to believe that the real markets may in many cases be larger.

As it stands, Germany's role looks preponderant. The national grid operators in Germany see already for 2015 a total installed PV capacity of 39 GW in the country – against 17.3 GW at the end of 2010. Following the roadmap of the German Solar Industry Association, the PV capacity could well attain 52–70 GW by 2020. In agreement with the competent Ministry of the Environment, the contribution of PV-generated electricity as part of total consumption could increase from 2% in the course of 2010 to 10% by that year.

Outside Europe, "Bloomberg New Energy Finance" sees a realistic goal of a 30 GW total capacity installed in the USA by 2020.

China originally aimed at 20 GW of Solar Electricity installed by 2020. But in the wake of the nuclear accident in Japan, the government is considering the option to raise the targets for 2015 from 5 to 10 GW and for 2020 to 50 GW [42].

Osamu Ikki of the RTS Corp. in Tokyo anticipates for Japan a cumulated PV capacity of 28 GW by 2020 [43].

In summary, the cumulated PV capacity to be installed and grid-connected globally by 2020 may become the following:

- Germany 52 GW
- China 50 GW
- EU without Germany 32 GW
- USA 30 GW
- Japan 28 GW

The sum of 192 GW for these leading countries may well be a minimum. And there is a high probability that PV will spread to many other nations worldwide. For instance, the analyst Ikki mentioned before sees on that horizon a total grid-connected PV capacity of 340 GW.

That would be almost 10 times the global capacity seen at the end of 2010 with 38 GW.

The *annual* PV market might reach 60 GW in 2020 – compared to 17 GW in 2010. For 2011 and 2012 analysts expect a global market volume of slightly over 20 GW. By 2030 it might increase to 200 GW, Ikki speculates [43].

Solar industry associations from Europe, the USA, and others presented to COP 16, the last Climate summit in Cancun in 2010, the report "Seizing the Solar Solution" [44]. The profession advocates a global capacity of 980 GW by 2020.

The International Energy Agency IEA in Paris, known rather for their conservative opinion in energy matters, speculates that by 2050 a quarter of the world's electricity needs could be covered by PV [45].

In summary, the years 2009/2010 have brought a turning point for the implementation of PV on a large scale. In particular in Europe, PV has now become a leader of all power investment. Across the board, no other type of power generation, not even wind power, saw the same level of investment, leaving alone nuclear energy that has been left far behind.

There are important consequences for global power generation in the longer run. Again in Europe, a large consensus has grown in the last few years that on the horizon of 2050, "green house gas" emissions must be reduced by at least 80%. The option of atomic power in this perspective is losing ground all the time, in particular after the accident in Fukushima. The only option that remains then for "carbon-free" power generation are the renewable energies. That reasoning does not even take into account that the conventional energies are finite resources and solar energy is not. As a consequence, it is essential to move in the next few decades to a 100% supply scenario with these "natural energies" that are all derived from the Sun.

In this situation, the good news is that PV becomes at last a credible contender. It is not unreasonable to believe that PV could eventually meet between one third and one half of all electricity demand in our countries. It is actually PV that opens a new door to a better energy world, one that is decentralized, with less need for high-voltage power lines than more, and higher energy self-sufficiency for everybody.

PV's promise used to be a dream – now it has become a very solid political and industrial business model. As mentioned here above, Germany has taken the driving seat in modern PV market implementation. There is a large political consensus in German society and across all political parties. Germany is happy with the experience it had thus far with the development of its PV markets – and the hardest is behind as grid parity in Germany is readily around the corner. And the German commitment for PV has a solid base as its economy, Europe's largest, is strong and booming.

The second pillar of PV's high credibility is the new PV industry that was built on hundreds of \$ billions of investment, a young and enthusiastic work force of over 200,000 people around the world, and last not least, a highly motivated R&D community.

## Bibliography

1. Wang J (2011) IMS Research. In: 6th AsiaSolar PV industry forum, Shanghai, 5–7 May 2011
2. Starr MR, Palz W (1983) Photovoltaic power for Europe, an assessment study. Reidel, Dordrecht. ISBN 90-277-1556-4
3. Neue Energie Berlin, 4 Apr 2011, p 55
4. Wrixon GT, Rooney A-ME, Palz W (1993) Renewable energy-2000. Springer, New York. ISBN 0-387-56882-4
5. Bundesverband Solarwirtschaft (2010) sonnenseite.com/Franz Alt: 1-1-2011
6. German Bundesnetzagentur quoted in guntherportfolio.com 3-23-2011
7. Neue Energie Berlin, 3 Mar 2011
8. RTS Corp. Tokyo (2011) PV activities in Japan and global PV highlights, vol 17, 3 Mar 2011
9. Lawrence Berkeley Nat. Lab. Dec 2010: PV installed cost trends
10. renewableenergyworld.com: 4-1-2011
11. Paula Mints (2010) electroiq.com: 10-22-2010
12. Bloomberg, VIP briefing Feb-2011 newenergyfinance.com
13. sonnenseite.com/Franz Alt: 1-17-2011
14. Systèmes Solaires Paris Nov/Dec 2010, p 65
15. renewableenergyworld.com Jan 2011
16. renewableenergyworld.com Feb 2011
17. renewableenergyworld.com March 11
18. Fell HJ, (2010) MP private communication
19. Letendre SE, The Grid Parity Fallacy, Renewable Energy World, 9-16-2010
20. Seltmann Th sonnenseite.com 1-30-11
21. enbausa.de 3-10-2011
22. enbausa.de 7-5-2010
23. pv-magazine.com 3-22-2011
24. pv-magazine.com 3-8-2011
25. newenergyworldnetwork.com 3-21-11
26. thesolarfuture.com March 2011
27. greentechmedia 3-10-2011
28. greentechmedia 3-4-2011
29. renewableenergyworld.com 7-5-2010
30. reuters.com 3-3-2011
31. RENEWABLES 2011, draft Global Status Report, REN 21, Paris, copyright GTZ/GIZ, Germany
32. solarnovus.com 3-8-2011
33. Haiyan QIN (2011) DG China General Certification Center, Beijing May 2011, personal communication
34. Solarbuzz, cited by pv-tech.org Mar 2011, and RTS Corp. Tokyo PV activities in Japan, vol 17, 4 Apr 2011

35. electroiq.com 2/23 2011
36. Palz W (1978) Solar electricity. UNESCO, Paris. ISBN 92-3-101427-7
37. Böer KW (2011) p-type emitters covered with a thin CdS layer show a substantial improvement of  $V_{oc}$  and FF. Sol Energy Mater Sol Cells 95:786
38. Palz W et al (2010) Power for the world. Pan Stanford, Singapore. ISBN 978-981-4303-37-8
39. solarbuzz.com, 4-25-2011
40. European Environment Agency eea.europa.eu
41. ECN, Petten, NL, ecn.nl
42. Sicheng W (2011) Energy Res. Inst. Beijing. In: 6th AsiaSolar PV industry forum, Shanghai, 5-7-2011
43. RTS Corp. Tokyo PV activities in Japan, vol 16, 12 Dec 2010
44. SolarCOP16 seia.org
45. reuters.com 4-19-2011 © Wolfgang Palz, Brussels, May 2011