Journal of Intelligent Manufacturing (2019) 30:2307–2326
https://doi.org/10.1007/s10845-018-1392-0

Materials informatics

Seeram Ramakrishna[1] · Tong-Yi Zhang[2] · Wen-Cong Lu[2] · Quan Qian[2] · Jonathan Sze Choong Low[3] · Jeremy Heiarii Ronald Yune[3] · Daren Zong Loong Tan[3] · Stéphane Bressan[4] · Stefano Sanvito[5] · Surya R. Kalidindi[6]

Received: 5 September 2017 / Accepted: 9 January 2018 / Published online: 17 January 2018

Abstract
Materials informatics employs techniques, tools, and theories drawn from the emerging fields of data science, internet, computer science and engineering, and digital technologies to the materials science and engineering to accelerate materials, products and manufacturing innovations. Manufacturing is transforming into shorter design cycles, mass customization, on-demand production, and sustainable products. Additive manufacturing or 3D printing is a popular example of such a trend. However, the success of this manufacturing transformation is critically dependent on the availability of suitable materials and of data on invertible processing–structure–property–performance life cycle linkages of materials. Experience suggests that the material development cycle, i.e. the time to develop and deploy new material, generally exceeds the product design and development cycle. Hence, there is a need to accelerate materials innovation in order to keep up with product and manufacturing innovations. This is a major challenge considering the hundreds of thousands of materials and processes, and the huge amount of data on microstructure, composition, properties, and functional, environmental, and economic performance of materials. Moreover, the data sharing culture among the materials community is sparse. Materials informatics is key to the necessary transformation in product design and manufacturing. Through the association of material and information sciences, the emerging field of materials informatics proposes to computationally mine and analyze large ensembles of experimental and modeling datasets efficiently and cost effectively and to deliver core materials knowledge in user-friendly ways to the designers of materials and products, and to the manufacturers. This paper reviews the various developments in materials informatics and how it facilitates materials innovation by way of specific examples.

Keywords Materials informatics · Materials data analytics · Materials modelling · Materials data mining · Materials selection · Materials web platform · Materials 4.0

Introduction

Emerging technologies and products from smart phones, personal healthcare devices to electric cars are dependent on the development and availability of suitable materials (Phillips and Littlewood 2016). Materials play an enabling role in clean water, air, energy and environment, human health and wellbeing, smart living and transportation, and safety and security (Dima et al. 2016). Several types of improved materials are needed for innovative products, while not every application requires a radically new material. They include light-weight materials, energy materials, biomaterials, food packaging materials, nanocomposites, electronic materials, thermoelectric materials, polymers, metals, ceramics, soft materials, and materials for extreme environments (Dean 1990).

✉ Seeram Ramakrishna
seeram@nus.edu.sg

1 Department of Mechanical Engineering, National University of Singapore, Institution of Engineers Singapore, and SPRING, Singapore 119260, Singapore

2 Materials Genome Institute (MGI), Shanghai University (SHU), and Shanghai Materials Genome Institute, Shanghai 200444, China

3 Singapore Institute of Manufacturing Technology, ASTAR, Singapore, Singapore

4 School of Computing, National University of Singapore, Singapore, Singapore

5 School of Physics, AMBER and CRANN Institute, Trinity College, Dublin 2, Ireland

6 Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Drawing parallels to the human genome, the materials genome comprises all the elements in the periodic table (de Pablo et al. 2014). However, we do not yet fully understand how function emerges from assembling appropriate atoms in the right way as well as how to precisely make the designed material. Until recently, an extensive trial and error approach has been the main mode of developing new materials. Moreover, the data generated is often not shared with others in efficient and effective modes, thus leading to loss of information, underutilization of materials information, redundancy of experiments and longer cycles of materials development (Jain et al. 2016). One of the earlier and main challenges that have been tackled by the material community is the accessibility of data through the development of high-throughput computation and combinatorial experiments, where data sharing in shortening the innovation cycles has been recognized as a key enabler (Kalidindi and De Graef 2015; Puchala et al. 2016). Mathematical and statistical methods are needed to extract patterns that can be leveraged for material discovery, design and optimization (Rajan 2015).

The discovery and optimization of materials is time-consuming, labor intensive, complex and expensive. A new consumer product from invention to widespread adoption takes about 2–5 years, but doing the same for a new material takes about 15–20 years. There is a need for matching materials innovation with the accelerated pace of new product designs and shorter cycles of manufacturing (Kalidini et al. 2016; Panchal et al. 2013). Product designers need precise information on materials functional performance and environmental impact. With the availability of cheaper sensors and faster communication technologies, vast amounts of data on in situ service performance of the products and materials can be easily collected. Emerging information, computation, communication technologies which include big data analytics, algorithms, data mining, artificial intelligence (AI), machine learning, industrial internet or internet of things (IIoT), high performance computing and cloud computing are presenting unique opportunities to materials design and materials selection based on the functional performance requirements. They can be employed to analyze larger and larger volumes of materials related data collected via ubiquitous sensors and industrial internet of things, and created by combinatorial materials science, high throughput analytical techniques, computational materials science based on accurate electronic structure methods, and materials property and failure prediction software and tools (Kalidindi et al. 2010).

In the footsteps of bioinformatics and the successful acceleration of genomics, material science and engineering can benefit from a large scale, collaborative and interdisciplinary approach to the design, development, selection and application of materials and processes leveraging the developing information, computation and communication technologies. We refer to such a holistic approach as materials informatics (Rajan and Seeram 2018). Materials genome, real time materials informatics, hierarchical materials informatics, web-based materials data sharing, etc. are other names appearing in the literature. Mulholland and Paradiso (2016) described it as algorithmically analyzing materials data across the product life cycle, i.e. selection, manufacturing and certification with a focus on reducing time to market for new advanced materials technologies. Rodgers and Cebon (2006) and Wang et al. (2014) described it as the application of computational methodologies to processing and interpreting scientific and engineering data concerning materials. In our broader view, the materials informatics employs techniques, tools, and theories drawn from the emerging fields such as data science, internet, computer science and engineering, and digital technologies to the materials science and engineering to accelerate materials, products and manufacturing innovations.

On a limited scale, a few materials databases, precursors to the current day material informatics, already exist. Major impetus for the materials informatics came from the materials genome initiative (MGI) of USA launched in 2011. As a part of MGI initiative, the US National Science Foundation, Department of Defense, Department of Energy, and National Institute of Standards and Technology (NIST), funded several individual projects as well as large scale centers. The overarching goal is to accelerate the discovery and development of new and improved materials at a fraction of the cost (National Science and Technology Council 2011; McDowell and Kalidindi 2016). The MGI infrastructure platform encompasses computational tools, experimental tools, and digital data. Accelerated materials design and deployment is envisaged by: (1) developing effective and reliable computational methods and software tools, (2) developing high-throughput experimental methodologies to validate theories and to provide reliable experimental data to the materials databases, and (3) establishing reliable and widely applicable databases and materials informatics tools.

The goal of this emerging field is to achieve high-speed and robust acquisition, management, multi-factor analyses, and dissemination of diverse materials data. This encompasses more specialized earlier developments such as materials property databases, combinatorial materials synthesis, materials data management, process modeling, life cycle inventory, life cycle impact assessment, and product life cycle management. Proponents of materials informatics see its analogy to the bio-informatics which applies software tools for understanding the biological data. It led to the personalized medicine based on the deciphered human genome. Moreover, the data storage technologies are highly developed. For example, the European Bioinformatics Institute stores approximately 20 petabytes of data, and the European Organization for Nuclear Research (CERN) stores over

200 petabytes. Hence, the materials informatics community believes that it has the potential to provide deeper understanding and derive greater insights by applying lessons learned from data gathered on one type of material to others. They see the benefits of materials informatics in terms of accelerated insertion of materials, thereby saving millions of dollars otherwise needed for the conventional way of materials development. It breaks down the barriers between data management, quality standards, data mining, exchange, and storage and analysis, as a means of accelerating scientific research in materials science.

The structure of this review paper is as follows. Following the Introduction, second section briefly describes web-based materials informatics platforms that are continuously and quickly growing-up to form the solid ground of the newly emerging materials informatics field. Materials informatics integrates materials science and engineering and artificial intelligence with the hub of machine learning and big data. Machine learning is an additional tool to the global materials community. Third section concisely introduces data-driven materials design by using machine learning approaches, where machine learning methods are tersely described in first subsection, descriptor selection in second subsection, the newly developed technique of adaptive material design in third subsection, and model explainability of machine learning in fourth subsection. The state-of-the-art applications of material informatics are overviewed in fourth section. Future directions of material informatics are briefly discussed in fifth section. Finally, sixth section gives the conclusions.

Web-based materials informatics platforms

Materials informatics trend is web-based platforms and data infrastructure with which it is easy to search for specific materials information from anywhere in the world and anytime. The materials informatics, mainly from the academic institutions and governmental organizations, are free to use; while the commercial ones, from professional societies and private companies, are pay for a license and pay per usage. Various facets of materials science and engineering involves generating reliable data and information during materials synthesis, processing, modeling and characterization at various length scales, materials changes and performance during the fabrication of products and service life of the product, and consequences of action taken at the end-of-life of product. These facets of materials science and engineering lead to high volumes and heterogeneity of data. Moreover, the data comes from diverse sources in different formats. The web-based materials informatics platforms to support data generation, acquisition, storage, mining, processing, management, and visualization services. The multi-faceted nature of materials informatics is illustrated in the schematic shown in Fig. 1. Table 1 lists the various materials informatics efforts around the world. The sections to follow will describe the details of these efforts.

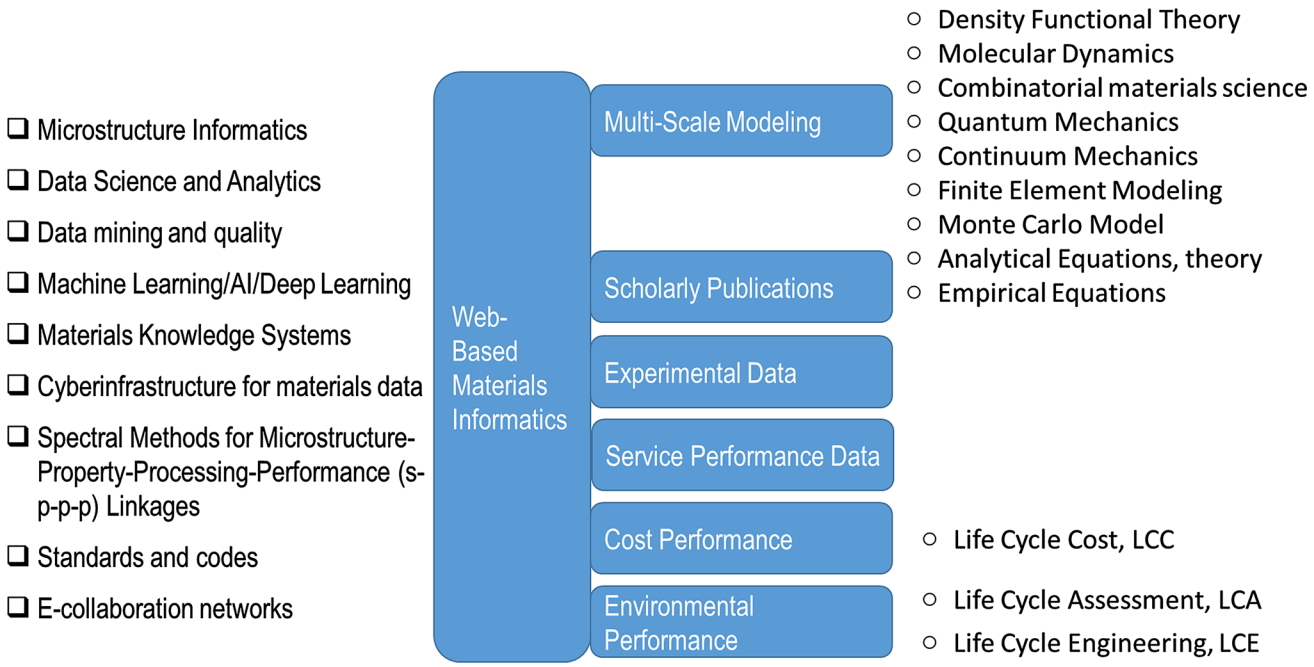

Fig. 1 Web-based materials informatics platforms are built on data infrastructure and analytic tools stated on the left side of the schematic. Data is generated and fed from the tools and means indicated on the right side of the schematic

Table 1 Materials informatics efforts around the world (Kalidindi et al. 2016a, b)

Materials informatics (compatible with Windows, Mac OS X, and Linux)	Host
AFLOWLIB	AFLOW consortium, Duke University
ASM medical materials database	ASM International
CALPHAD (calculation of phase diagrams) for metallic alloys	National Institute of Standards and Technology, NIST
CHiMaD	NIST (National Institute of Standards and Technology) Center for Hierarchical materials design (CHiMaD) at Northwestern University, Argonne National Laboratory and The University of Chicago
Citrine informatics—thermoelectric materials	US based private repository for materials data spanning from development to manufacturing
Granta material intelligence	A private company at Cambridge, UK
Harvard clean energy project on organics for photovoltaic and electronics	http://cleanenergy.molecularspace.org/
Materials common-based on phase equilibria, microstructures and mechanical properties of metals and alloys	Department of Energy funded PRISMS (Predictive Integrated Structural Materials Science) Center at the University of Michigan, USA
Materials project (MP)	Lawrence Berkeley National Laboratory
MatNavi—metals, polymers, ceramics	National Institute of Material Science of Japan (NIMS Japan)
NanoMine for nanocomposites	Northwestern University
Novel materials discovery (NoMaD)	Europe http://nomad-repository.eu/cms/
Open quantum materials database (OQMD)	Northwestern University
OpenKIM project on interatomic potentials	University of Minnesota
www.matweb.com	Web based materials property platform

Since 2011, a number of projects for data curation have been initiated. The Materials Project at Lawrence Berkeley National Laboratory, the OpenKIM project on interatomic potentials at the University of Minnesota, the NIST (National Institute of Standards and Technology) Center for Hierarchical Materials Design (CHiMaD) at Northwestern University, Argonne National Laboratory and The University of Chicago, and the materials common at the University of Michigan PRISMS (Predictive Integrated Structural Materials Science) Center funded by the Department of Energy are to name a few examples. Efforts are also made to develop protocols for annotation of the materials data, data structures, archiving data on the web such that the value of the data is maintained over time, and the web-based database remains available for reuse and preservation. For example, the National Institute of Standards and Technology (NIST) facilitated efforts on data curation by developing a phase-based materials ontology for ensuring consistency among disparate phase-based materials community (NIST 2013). They adopted unified modeling language (UML) and XML schema for this purpose. The database infrastructure is based on NoSQL as well as other standard relational technologies. They also adopted traditional APIs, Web APIs (REST), and data exchange facilities and formats (XML, JSON, BSON) to provide flexible data access to the web-based databases, scalability and access to new tools of big data analytics, machine learning and artificial intelligence.

Novel materials discovery, NoMaD is a European materials database. The National Institute of Material Science of Japan (NIMS Japan) maintains MatNavi database. MatWeb is another searchable online material database with data cataloged from several manufacturers and suppliers. The Materials Intelligence system from Granta Design, Cambridge, UK is a commercial platform that integrates materials data with a variety of software tools and provides Ashby charts for different types of materials. The US based Citrine Informatics offers private repositories for materials data spanning from development to manufacturing. The Citrine Informatics created a web-based database using a machine-learning-based recommendation engine for identifying new thermoelectric materials using a large body of experimental thermoelectric characterization data and first-principles-derived electronic structure data as the training set. In other words, it provides datasets collated from multiple sources and data-driven material design tools.

National Institute of Standards and Technology (NIST) has developed phase diagram based CALPHAD (CALculation of PHAse Diagrams) for metallic alloy materials as a part of MGI (Kaufman and Ågren 2014). It captures both experimental and computational data related to thermodynamics, kinetics diffusion, molar volume, elastic properties, electrical

conductivity, thermal conductivity and interfacial energies. In recent years, it is being developed into OpenCalphad (OC), an informal international collaboration of scientists and researchers interested in the development of high quality software and databases for thermodynamic calculations using the CALPHAD. The materials common by PRISM at the University of Michigan focuses on phase equilibria-microstructures-mechanical properties of metals and alloys (Puchala et al. 2016). Materials Commons is being developed as platform for use by the global materials community for accelerating the prediction of materials phenomena. Other databases include Citrine Informatics' system, NanoHuB, the National Data Service's Materials Data Facility, and the Crystallographic Open Database (COD) integrated with computational ICSD.

The open quantum materials database (OQMD) and the Materials Project (MP) utilize results from density functional theory computational methods coupled with data analytics for screening and discovery of promising material systems with enhanced properties for batteries and catalysts. NanoMine developed by the researchers at the Northwestern University focuses on nanocomposites materials informatics in order to facilitate efficient material selection and design. AFLOWLIB is a large database (about 1,6M entries to date) of electronic structure calculations performed at the level of density functional theory. It is maintained by Duke University and contains data produced by a consortium of 14 research groups distributed over three continents. AFLOWLIB contains data for about 70% of the compounds whose crystal structures are reported in the ICSD repository, in addition to several libraries of hypothetical new phases, including binary and ternary intermetallic, Heusler alloys, etc.

Diverse stakeholders of nascent materials informatics field such as the private entities (Citrine Informatics and Springer Nature Nano), non-profit societies (ASM International, ASME, ASTM), and government agencies (NIST) are shaping standards for open data frameworks, systems, and ontologies that flexibly accommodate data for purposes of record keeping, easy retrieval via web searches, and analysis. In order to encourage progress via open competition, NIST and Citrine Informatics hosted Materials Data Challenge and Materials Hackathons respectively.

It is to be noted that the aforementioned materials informatics infrastructure are in early stages of development and not yet reached the stage of easily and flexibly meet the needs of a variety of users and products.

Table 2 provides information on various software approaches for data formats, data security, data dissemination strategies, data analysis and visualization. They are helpful to curate, share and mine materials data from scientific literature and diverse sources for further insights. Currently the data formats used in materials science and engineering are very diverse and hence, pose challenges to creating easy to use databases.

Table 3 lists various software that have been widely used for modeling and simulation to predict materials properties. Life cycle assessment and life cycle engineering related software and web-based databases are listed in Table 4. These lists are not meant to be exhaustive but provide a sampling of global efforts in this direction.

Machine learning

As indicated in Fig. 1, machine learning and data mining play an essential role in emerging materials informatics. Machine learning is developed along with the development of computer science and now becomes an important subfield of computer science. Machine learning gives computers the ability to learn from data and make predictions based on data (Samuel 1967). Mitchell (1997) defined machine learning as "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." In materials informatics, machine learning algorithms learn from existing material data, which include input information represented by descriptors and output responses that are usually material properties or/and performance of interest, in order to design and discover new materials with improved targeted properties or/and performance. Since material data are usually sparse, adaptive design with feedback from experiments or/and computations has been proposed to enhance the ability of machine learning. After the introduction of machine learning algorithms, the descriptor selection, adaptive material design, and machine learning model explainability are described in reader-friendly manner with successful cases of materials machine learning approaches.

Machine learning algorithms for materials informatics

Regarding learning style (Brownlee 2013), there are three categories in machine learning algorithms, called supervised learning (Decision Tree, Boosting, Artificial Neural Network, Support Vector Machine, etc.), unsupervised learning (Clustering, Associate Rules, etc.), and semi-supervised learning (Friedman 2001). Supervised learning uses labeled data to train a machine learning model, where the data is called training data and the trained model will have the capability to predict the relationship between targeted properties and features. On the other hand, unsupervised learning clusters only unlabeled data. In semi-supervised learning, unlabeled data are used along with the labeled data to improve the accuracy of the models on the training data.

Table 2 Software tools and methods for management of entire data life cycle (Dima et al. 2016; Jain et al. 2016; Le and Winkler 2016; Puchala et al. 2016).

Purpose	Tool or method	Description
Data verification, data formatting	• ESTEST framework for validation and verification of electronic structure codes—Qbox, Quantum Espresso, Siesta, ABINIT, and The Exciting Code	To ensure that the data is sensible and relevant, information is verified and formatted according to specifications. For data exchange to happen, different software may require specific formats before they can be interpreted properly
	• Comma-Separated-Values (CSV) format for tabular data	
	• HDF5 and netCDF formats for complex datatypes	
	• Yet Another Markup Language (YAML), Extensible Markup Language (XML) and JavaScript Object Notation (JSON) for structured meta data	
	• ChemML and MatML are for more scientific specifications	
	• UrXML to format XML in non-XML documents	
	• Materials Information File (MIF) as a structured material data format	
	• Binary Large Object (BLOBs) to store large data objects as binary or character data	
Data curation, data security	• SQL for table structured databases—MySQL	As data is collected and represented in a database management system, ongoing maintenance of data is necessary. This includes ensuring the authenticity and integrity of data through proper data security measures
	• NoSQL for document-oriented databases—MongoDB,	
	• Hypertext Transfer Protocol Secure (HTTPS) for secure data communication over the web	
	• Antivirus protection to prevent data-stealing—McAfee, Symantec, Kaspersky	
	• Encryption for authentication and authorization—BitLocker	
	• Virtual Private Network (VPN) for secure connection in organizations	
	• Data backup solutions to prevent data loss—Acronis, Backblaze, IBM Backup	
Data sharing, data centralization, data dissemination	• ZIP archive for compressed file download	Technical aspects that allow for efficient data transfers, coupled with repositories where data are predominantly collated and tools that encourage collaboration for proper dissemination of shared data
	• REpresentational State Transfer (REST) or RESTful web services for client-server resources exchange	
	• Hypertext Transfer Protocol (HTTP) as the foundation for data communication over the web	
	• Uniform Resource Locator (URL) as web address to access web resources	
	• NGINX for resource efficient web services	
	• Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)	
	• Source-code repositories—Git, Mercurial, Subversion	
	• E-collaboration tools—Jupyter, Galaxy, Pegasus, KNIME, Orange, and gUSE	
	• Collaborative platforms—Materials Data Curation System (MDCS); NIST Materials Resource Registry (NMRR); Materials Commons; Citrine Informatics	
	• ChemSpider search engine for the Chemistry community	
	• DSpace open source repository application	

Table 2 continued

Purpose	Tool or method	Description
Web interface, computational algorithms, data visualization	• Frameworks for web development such as Django, Flask, Pyramid, Laravel, Ruby on Rails	Existing frameworks that help in deploying web-based interfaces for user interaction. Computation methods and algorithms are introduced to transform data into valuable information which can be further illustrated through visual representation of data
	• E-science gateway—MATIN based on HUBzero	
	• User interaction via a graphical user interface (GUI) or a representational state transfer application programming interface (REST API)	
	• MATLAB Neural Network Toolbox	
	• Rotation forest algorithm	
	• Evolutionary algorithms—SourceForge genetic algorithms; HeuristicLab; AMALGAM; NSGA-II; RosettaCode; GPLAB	
	• Data analysis tools—R, SciPy, NumPy, Scikit-learn, StatsModels, Pandas, MATLAB	
	• Data visualization tools—Chart.js, Tableau, Visual.ly, D3.js, Google Charts, CartoDB	

Table 3 Materials modeling and simulation software

Common names of software
DS Solid Works
LS-Dyna (Ansys)
Autodesk-ALGOR and PLASSOTECH
COSMOSWorks
SpaceClaim 3D
ABAQUS (Simulia)
Matlab (MAthWorks)
Pam-Crash (ESI)
Mathematica
LAMMPS
DICTRA
MICRESS
Python-based materials knowledge systems (PyMKS)

Table 4 Life cycle assessment software and databases

Software and databases	Provider
Boustead Model 5	Boustead Consultants (www.boustead-consulting.co.uk)
Carbon Calculator	Carbon Trust, London (www.carbontrust.com)
CES Eco'12	Granta Design, Cambridge, UK (www.grantadesign.com)
EarthSmart	www.earthshiftglobal.com
Eiloca	Carnegie Mellon Green Design Institute, USA (www.eiloca.net)
GaBi-LCA software system	PE International, Germany (www.gabi-software.com)
GHG Protocol Organization	http://www.ghgprotocol.org/Third-Party-Databases
GREET	US Department of Transport (www.transportation.anl.gov)
KCL-Eco 3.0	KCL, Finland (www.kcl.fi)
LCA Calculator	IDC, London, UK (www.lcalculator.com)
MIPS	Wuppertal Institute (www.wupperinst.org)
Okala Ecodesign Guide	Industrial Design Society of America (www.idsa.org/okala-ecodesign-guide)
OpenLCA	GreenDelta (www.openlca.org)
SimaPro-LCA software system	PRé Consultants (www.pre.nl), Netherlands https://simapro.com/
Team (EcoBilan)	PricewaterhouseCoopers (www.ecobalance.com)

According to the format similarity or the function similarity, machine learning algorithms are divided into eight groups: dimensionality reduction, regression, decision tree, Bayesian algorithm, clustering, artificial neural network, deep learning, and ensemble, in which the four groups of dimension reduction, regression, classification, and clustering are commonly used in materials informatics. Dimension reduction includes initial and final feature selections, and feature transformation. Feature selection aims to reduce features in high-dimensional space to low-dimensional space and to find a subset, which influence mostly to the targeted properties, from the initial feature set or/and to rank the initial features based on the influence. Regression and classification algorithms are usually used for macro or micro-level material properties predictions. Clustering is used as outlier detection. More attention should be paid to outliers because they might be caused by noises or new inventions. Popu-

Table 5 Popular machine learning algorithms used in materials informatics

Algorithm	Category	Popular algorithms
Dimensionality reduction	Dimensionality reduction	Principal Component Analysis (PCA), Principal Component Regression (PCR), Partial Least Squares Regression (PLSR), Sammon Mapping, Multidimensional Scaling (MDS), Projection Pursuit, Linear Discriminant Analysis (LDA), Mixture Discriminant Analysis (MDA), Quadratic Discriminant Analysis (QDA), Flexible Discriminant Analysis (FDA)
Regression	Regression	Ordinary Least Squares Regression (OLSR), Linear Regression, Logistic Regression, Stepwise Regression, Multivariate Adaptive Regression Splines (MARS), Locally Estimated Scatterplot Smoothing (LOESS), Ridge Regression, Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net, Least-Angle Regression (LARS)
Decision tree	Regression and classification	Classification and Regression Tree (CART), Iterative Dichotomiser 3 (ID3), C4.5 and C5.0 (different versions of a powerful approach), Chi-squared Automatic Interaction Detection (CHAID), Decision Stump, M5, Conditional Decision Trees
Bayesian	Regression and Classification	Naive Bayes, Gaussian Naive Bayes, Multinomial Naive Bayes, Averaged One-Dependence Estimators (AODE), Bayesian Belief Network (BBN), Bayesian Network (BN)
Artificial neural network	Regression and classification	Perceptron, Back-Propagation, Hopfield Network, Radial Basis Function Network (RBFN)
Deep learning	Regression and classification	Deep Boltzmann Machine (DBM), Deep Belief Networks (DBN), Convolutional Neural Network (CNN), Stacked Auto-Encoders
Clustering	Clustering	k-Means, k-Medians, Expectation Maximization (EM), Hierarchical Clustering
Ensemble	Regression and classification	Boosting, Bootstrapped Aggregation (Bagging), AdaBoost, Stacked Generalization (blending), Gradient Boosting Machines (GBM), Gradient Boosted Regression Trees (GBRT), Random Forest

lar machine learning algorithms in materials informatics are described in Table 5.

It might be a great challenge in machine learning how to choose an optimal algorithm with appropriate fitting of available data without overfitting or under fitting. For example, Raccuglia et al. (2016) conducted five machine learning models of decision tree (C4.5), random forest (size 100 and 1000), logistic regression, k-nearest neighbors ($k = 1, 2$, and 3) and support vector machine (SVM) (Vapnik 1995, 1998) on 3955 chemical reaction with 273 descriptors per reaction. Their results indicated that the SVM model yielded the highest classification accuracy (74.1%) over the others based on 15-cross-validation tests (Efron 1983).

Descriptor selection

In machine learning, descriptors are called features or attributors as well. Many attributors of atoms and electrons are widely used as descriptors in materials informatics, especially in high-throughput computation based materials informatics. Actually there are multiple types of descriptors. In chemical reaction based material synthesis, reactants and reaction conditions are naturally descriptors as inputs and the responses are products of the chemical reaction. In metal forming, descriptors could be raw metal composition and microstructure, the shape and size of the raw metal, and the metal forming conditions. In any circumstances, descriptor selection is one of the most important steps in materials data mining. Descriptors might be regarded the decisive factors in the construction of mechanism-independent correlations between the targeted properties and descriptors, which will become guidance for materials design and discovery. In mechanism-based analysis of data, descriptors are well defined and targeted properties are explicitly or implicitly expressed in terms of descriptors, and the expressions appear in analytic form or others. Descriptors are called variables in mechanism-based formulation and modeling, where it is very much straightforward to identify whether these variables are independent variables or not. In mechanism-independent data mining, the situation is completely different from the mechanism-based data analysis. Data mining deals with mechanism unclear problems, where only data are available. The number of available materials data is much smaller than the number of the virtual data in the search space. Therefore, appropriate descriptors with a statistic model (regressor) might require less training data and render the trained statistic model more power in prediction. The purpose of descriptor selection in materials data mining is to find the most influential features for modeling of targeted properties without redundancy so that descriptor selection should be carried out in such a way that the dimensionality of input space can be reduced without loss of important information.

Ghiringhelli et al. (2014) proposed four important properties of a descriptor. (1) A descriptor uniquely characterizes a material as well as property-relevant elementary processes. (2) Materials that are very different (similar) should be characterized by very different (similar) descriptor values. (3) The determination of the descriptor must not involve cal-

culations as intensive as those needed for the evaluation of the property to be predicted. (4) The number of descriptors should be as low as possible (for a certain accuracy request). To demonstrate how meaningful descriptors can be found systematically, they took the energy difference between zinc blende semiconductors and wurtzite and rocksalt semiconductors as targeted property, where the energies of these semiconductors were calculated ab initially. The goal of this machine learning is to find descriptors that are able to classify distinctly the zinc blende semiconductors from the wurtzite and rocksalt semiconductors. For this purpose, few essential atomic features were combined in physical sense manner to form sums and absolute differences of homogeneous quantities. The combined features were combined further and further. Finally about 4500 feature candidates were formed and each of the about 4500 feature candidates had still an explicit expression of the essential atomic features. Then, the least absolute shrinkage and selection operator (Tibshirani 1996) was employed to select features from the about 4500 candidates, leading to the best (i.e., the lowest RMSE, see "Appendix 1") 1D, 2D and 3D features.

It should be emphasized again that domain knowledge must be fully utilized in the initial selection of descriptors. As described above, it might be good to select initial descriptor candidates as many as possible in order to avoid any potential missing of some key descriptors or/and to find the right format of descriptors. Initial descriptor candidates should be carefully examined to check the correlations among them. A Pearson correlation map is very often used to illustrate the correlations among initial descriptor candidates. The Pearson correlation map does not involve any regressors to analyze correlations of descriptors. Another technique to analyze correlations of descriptors without involving any regressor is called minimal-Redundancy-Maximal-Relevance (mRMR) (Peng et al. 2005) and has also great potential in materials informatics.

Xue et al. (2017) used only three features to predict the martensitic transformation temperatures of 1,652,470 compositions from initial 53 data. The three features were selected from initial 16 feature candidates via two steps. First, the Pearson correlation map was utilized to find the correlations among these 16 initial feature candidates by using the 53 training data. The Pearson correlation map indicates two highly correlated groups in the initial 16 feature candidates. Based on the Pearson correlation map and domain knowledge, Xue et al. (2017) selected 7 features from the initial 16 feature candidates. Second, the 7 features were further selected by the so-called subset selection method via a linear regressor of phase transition temperature versus descriptor(s). The mean squared error (MSE) and R^2 statistics (see "Appendix 1") were used to measure the fitting accuracy for each regression model as a function of the number of features with possible subset combination of descriptors used to fit the model. The subset selection method selected the final three features. Obviously, the subset selection method involves a regressor and thereby the final selected features might depend on the regressor involved. It is very popular that feature selection is combined with regressor.

Xiong et al. (2015) combined the genetic algorithm (GA) (Gen and Cheng 1997) with a support vector regression (SVR) model to select descriptors in the study of desired basal spacing of layered double hydroxides (LDH). The GA-SVR method selected the subset of descriptors and trained the model that was evaluated by the leave one out cross validation (LOOCV) simultaneously. The minimal root mean square error (RMSE) completed the training of the model and selected 4 descriptors from the 19 candidates. Then the machine learning gives the linear relationship between the basal spacing (d) of LDH compounds with the four descriptors and Xiong et al. (2015) verified this prediction by synthesizing and characterizing a new Mg-Al-CO_3 LDH compound.

de Jong et al. (2016) conducted data mining on the Voigt–Reuss–Hill (VRH) averages (see "Appendix 2") of elastic bulk modulus K and elastic shear modulus G of 1940 inorganic compounds, where data were obtained from first-principles quantum mechanical calculations based on Density Functional Theory (DFT). Based on the material knowledge at the atomic level, they initially selected 17 descriptor candidates including 8 composition descriptors and 9 structural descriptors for each compound. Then, de Jong et al. (2016) used a statistical learning technique, called the gradient boosting machine local polynomial regression (GBM-Locfit), to select the most important descriptors with Hölder means. The final GBM-Locfit model trained with the 1940 data determines four most useful descriptors for predicting bulk modulus K and shear moduli G, although the order of the four most useful descriptors in the prediction of bulk modulus K differs from that in the prediction of shear modulus G. It is very impressive that none of the models with more than four descriptors had significantly better predictive accuracy than these four-descriptors models, concluded from the comparison of prediction MSE and their associated standard errors. On the other hand, all other models with less than four descriptors had significantly less predictive accuracy.

Adaptive design in materials informatics

In materials informatics, one great challenge is that available dataset is usually small, while the search space of virtual dataset is large. A data mining algorithm or model is usually trained with a small dataset, called training dataset, of heterogeneously distributed data. Then, applying the trained model to the search space of huge numbers of virtual data may yield two distinct regions of exploration and exploitation in the search space. Caution must be exercised in designing

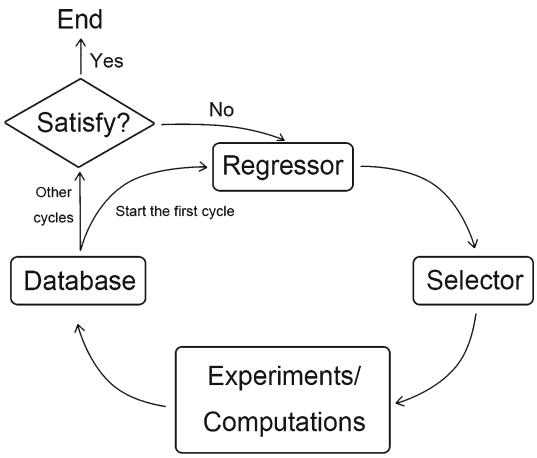

Fig. 2 Adaptive design approach integrating database, which is built based on domain knowledge and includes material data of targeted properties and descriptors, regressor, which is a statistics-based algorithm to regress targeted properties versus descriptors, selector, which is also a statistics-based algorithm to design new experiments or/and computations by optimizing the selection from exploration and exploitation regions, and feedback experiments/computations

next experiment or/and computation with machine learning results and the exploration and exploitation regions must be considered in the design. Xue et al. (2016) developed an adaptive design approach, as schematically shown in Fig. 2, to address such a challenge of data sparseness and their results show the success. The adaptive design approach adopts iterative feedback that incorporates data mining, design, and experiment (or/and computation). The adaptive design approach starts with initial data, combines data mining, experimental (or/and computational) design, and experiment (or/and computation), and feeds the new experiment (or/and computation) results back into the dataset. The iteration continues until the targeted properties are achieved. The success of Xue et al. (2016) indicates that adaptive design approach combining machine learning and experiment (or/and computation) is able to accelerate new materials discovery with targeted properties.

In the adaptive design approach, a database of interested materials with targeted properties to be pursued is an essential and key ingredient. Reliability tests and descriptor selection are usually conducted in order to build-up an initial dataset. Each data in the initial dataset contains the experimentally measured (or/and numerically computed) material information of properties (or/and performance) and descriptors (or features called here). Careful examining all data in the initial dataset is able to determine an appropriate size of search space. If the search space is too small, the targeted properties might be out of the research space, thereby increasing the bias error. On the other hand, a too large search space will make the estimation variance of machine learning results big. Alternately, one may choose the search space and then generate the initial dataset by designed experiments or/and computations. A machine learning algorithm, as a core ingredient in the adaptive design loop, is based on a statistical inference model and trained with such an initial training dataset. The algorithm is called the regressor here. The regressor after training builds up the mapping relationship with uncertainties between properties and features and after that the regressor is often named as the trained model. In general, the trained model is verified by cross-validation using various data subsets (slits). In the adaptive design approach, especially when the initial dataset is small, the trained model is directly applied to the search space of a virtual dataset. Except of the initial dataset, the properties have not been measured in the virtual dataset, but predicted by the trained model. Based on the initial dataset and the theoretical prediction from the regressor, a selector, as another key ingredient in the adaptive design loop, designs the next batch of experiments (or/and computations) by balancing the trade-off between exploitation and exploration regions. Xue et al. (2016) adopted few selectors including Min, Efficient Global Optimization (EGO), and Knowledge Gradient (KG) to recommend the next batch of experiments. Selector Min greedily recommends the regressor predicted material in the search space, i.e., a pure exploitation. The latter two selectors recommend new materials for next experiment by considering both exploration and exploitation. This is essentially the probability of improving the current best estimate of the targeted property by sampling estimates in the search space. The selector treats the uncertainties in the sampling estimates. The algorithm of selector EGO maximizes the 'expected improvement'. The EGO algorithm can automatically move onto regions of higher uncertainty after the local search space of lower uncertainty. Selector KG maximizes the same 'expected improvement' as that used in selector EGO with a minor change. The selector will determine, under the constraints of experimental/computational capability, how many experiments (or/and computations) will be conducted in the exploitation and exploration regions, respectively. The designed batch of experiments (or/and computations) is then conducted and the experimental (or/and computational) results will be collected and put back into the database. This is the end of the first cycle and also the starting of second cycle of iterations. Obviously, the dataset gets bigger after a cycle of iterations. The feedback from the selector designed experiments (or/and calculations) will definitely improve the subsequent machine learning model.

Balachandran et al. (2016) compared several adaptive design strategies, which were similar to the one adopted by Xue et al. (2016). The targeted properties were bulk, shear, and Young's moduli of 223 M_2AX compounds, which were obtained from first principles calculations. The ultimate goal was to discover materials with the targeted properties in as few cycles of iterations as possible. In addition to selectors of

EGO and KG, Balachandran et al. (2016) also used the following four selectors: (1) Max: Chooses the highest expected score from the regressor. (2) Max-A: Alternates between choosing the material with the highest expected score and the material with the most uncertain estimated score. (3) Max-P: Maximizes the probability that a material will be an improvement, without regard to the size of the improvement. (4) Random: Chooses randomly an unmeasured compound. Their results indicate that the KG design selector outperformed the purely exploitive selectors. There are always uncertainties in the material data. These uncertainties come from experimental errors (e.g., from processing conditions, inherent instrumentation limitations), computational errors, and the uncertainty of a regressor. The KG design selector is able to handle these uncertainties and thus provide the best design for next experiment or computation.

The prediction accuracy of a selector depends on the used regressor. In this sense, a combination of regressor:selector guides the next experiment or computation in the adaptive design. For example, Xue et al. (2016) investigated the performances of several regressor:selector combinations by cross-validation and based on the initial dataset, when regressor:selector combinations were trained on randomly chosen subsets from the initial dataset. Obviously, the best performed regressor:selector combination should be adopted in the recommendation of new experiments or/and computation.

In brevity, adaptive design suggests feedback from new experiment or/and computation which are recommended by the best performed regressor:selector combination, i.e., the best machine leaning predication. The new experiment or/and computation examine and verify the machine learning predication and the feedback of new experiment or/and computation results increases the database size dynamically with the cycle number of iterations. Then, the used algorithms of machine learning are trained again with the expanded database in order to recommend next batch of experiments or/and computations. The iterations are going on until the targeted properties are finally achieved.

Model explainability of machine learning

When choosing a machine learning model, one usually faces such a trade-off between accurate black-box models and less accurate but easy-to-explain white-box models. Many practical applications of machine learning demand that an adopted model has the ability to explain why and how certain predictions are made. For example, comparing with those black-box learning models (i.e. artificial neural network, support vector machine, etc.), rule learning (Fürnkranz et al. 2012) (i.e. RIPPER) is a white-box model that can learn set of rules from the training data. Therefore, these two kinds of models might be combined in materials informatics so that while-box models are used to analyze the successful results generated from black-box models in order to explain the successful results.

Raccuglia et al. (2016) used decision tree to explain the SVM learning results. Although the used SVM model has the best generalization performance (Raccuglia et al. 2016), the SVM model is opaque and hard to interpret the predicted results since the kernel function adopted in the SVM model is nonlinear. Thus, a 'model of the model' of a decision tree (Quinlan 1993) was constructed to re-interpret the machine learning results of original SVM model. For the decision tree construction, all data were relabeled with the predicted outcomes of the SVM model. This meant that the entire set of features used in the SVM model was available in the decision tree construction, where the feature sequence adopted in if–then criteria was determined by using a C4.5 decision tree algorithm (implemented in WEKA 3.7) to model those predicted outcomes of successful results. It should be highlighted that chemical hypotheses to guide future experiments can be generated from the flow chart of decision tree of human-interpretable if–then criteria. The constructed decision tree explained the targeted compound type in either a poly-crystalline or single-crystal form of templated vanadium selenites recommended from the SVM model. Three hypotheses generated from the decision tree were about the formation of templated vanadium selenites, categorized by the molecular polarizability of the amine. The chart flow and hypotheses provide guidance to future experiment with respect to how to generate and construct building units and how to avoid the undesirable building units.

It should be emphasized that all reliable materials data either successful or failed ones are useful in the progress of developing novel materials, since successful and failed experiments will be both necessary in the determination of boundaries between success and failure. Machine learning treats data not only in a black box, as here the used SVM algorithm, but also can interpret data in a human-interpretable if–then criteria manner, as here the used decision tree algorithm. Thus, using machine learning in materials discovery will greatly enhance the successful rate and considerably improve the understanding of complex behaviors of materials as well.

Regarding the important aspects of machine learning mentioned above, the following three aspects deserve more emphasis for material informatics.

(I) Heterogeneous materials data integration. Materials data, generally speaking, has the typical characteristics of multi-sources and heterogeneity. These diverse materials big data, if collected and fused into open and accessible databases, will find applications in materials discovery, materials selection, materials failure analysis and life prediction. Materials science and engineering is very much concerned with the rela-

tionship among materials processes, microstructure, properties, and performance, called the processing–structure–property–performance (PSPP) relationship. Materials informatics tries to analyze materials data to decipher the PSPP relationship via forward and inverse models. The challenge lies in the materials data sparseness. Most materials data are not big enough and scatter greatly due to their characteristics of multi-sources and heterogeneity. The diverse heterogeneous materials data and databases must be integrated and there are various levels to do the integration. Jain et al. (2016) proposed three technical approaches in data integration, which are data formats, data dissemination strategies, and data centralization. Over the integrated materials data and databases, knowledge based integration might be carried out based on data semantics which depends on knowledge representation formalisms, such as semantic nets, frames, rules and ontologies. Currently, the ontology based integration is the main stream in the knowledge based integration.

(II) Online learning and dynamic learning. Online machine learning, opposed to traditional batch learning scenario that generates the best predictor by learning the entire training data, will perform the learning model incrementally in which data becomes available in a sequential order (Shalev-Shwartz 2011). In some circumstance, computational data or/and experimental data are sequentially generated. Thus, it is possible for machine learning algorithms to dynamically take the online data in order to improve the computation or/and experiment. The adaptive design approach takes the dynamic learning so that the size of a studied database grows continuously until the goal is achieved.

(III) Domain knowledge representation and utilization. In machine learning, the domain knowledge (i.e., the knowledge of materials science and engineering), is still the core since the essential of materials informatics is about transforming materials data and information into academic knowledge and industrial applications. High level domain knowledge guides each step in the process and reduces the gap between information and knowledge. A trustful discovery of new promising materials, identification of anomalies, and scientific advancement depends on the knowledge representation, retrieval, acquisition, management, knowledge level modeling, etc., in the adopted machine learning.

Applications of material informatics

The materials informatics approach has been employed for multiple purposes and is expected to open up more applications in the future. Some examples of these applications are listed in Fig. 3 and illustrated in the following sections and case studies.

Prediction of properties: design and discovery of novel materials

Agrawal et al. (2014) and Agrawal and Choudhary (2016) cited an example of steel fatigue prediction using Mat-Navi database of NIMS, Japan. Materials informatics approach yielded accurate results and gave confidence to further build models based on experimental data to connect processing and composition directly. About 4500 new stable inorganic compounds were identified using materials informatics approach and were later confirmed experimentally. The materials informatics was applied to optimize the microstructure of Fe-Ga alloy (Galfenol) to match the target elastic, plastic and magnetostrictive properties. As mentioned above, Raccuglia et al. (2016) applied machine-learning algorithms-enabled materials informatics to synthesize templated vanadium selenites. This work is significant as it can be extended to other promising target candidates like zeolites for gas adsorption purposes. In the past, the formation of such compounds is not fully understood and they were developed primarily on exploratory syntheses. Materials informatics approaches provide deeper insights of mobility, photovoltaic properties, gas

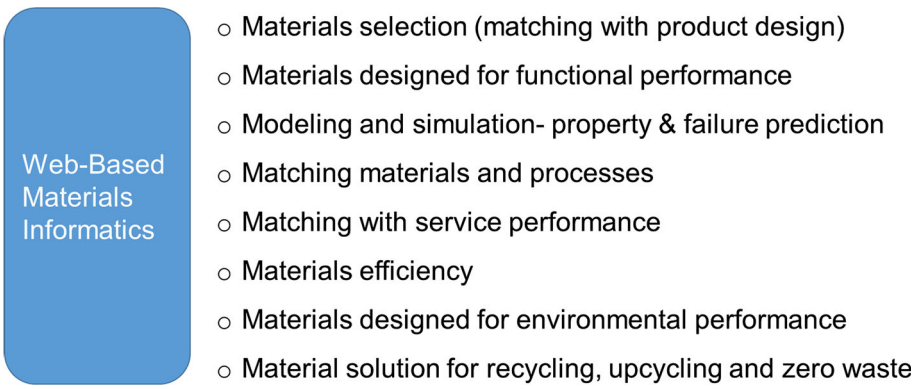

Fig. 3 Examples of applications of materials informatics

adsorption capacity or lithium-ion intercalation. Le and Winkler (2016) applied evolutionary algorithm to generate new catalysts, phosphors, and electroceramic materials. Lookman et al. (2016) applied materials informatics approach to discover new NiTi-based alloys $Ti_{50}Ni_{46:7}Cu_{0:8}Pd_{0:2}Fe_{2:3}$ with the smallest thermal dissipation. They used statistical inference and adaptive design for materials discovery with steps (a) assembling a library of crystal structures and chemistries and (b) defining the training space with a given number of samples and features, and (c) building an inference model using off-the-shelf pattern recognition tools, such as classifiers and regressors based on linear or kernel ridge regression, least squares regression, decision trees, Gaussian process modeling or support vectors (Zhao et al. 2003).

Importantly, in particular in cases where the data about materials have computational origin, the same dataset can be shared and used to identify novel compounds potentially relevant for completely different application spaces. For instance, the AFLOW consortium has constructed a large library of Heusler alloys (a family of ternary X_2YZ, such as Cu_2MnSn), comprising all the possible chemical compositions obtainable with 52 elements in the periodic table. This totals approximately 300,000 prototypes. The same dataset was then used to search for low-thermal conductivity semiconductors (Carrete et al. 2014) and for novel magnets (Sanvito 2017). In the first case an initial low-throughput set of calculations was performed to extract the lattice thermal conductivity of a small selected group of 32 compounds. Then in a second stage a machine learning method was applied to a second group of 450 alloys using the 32 as training set. Such second set was obtained by using mechanical stability and other criteria from an original pool of about 80,000 compounds (only half-Heuslers were considered). The method finally identified a number of low-thermal conductivity alloys and an empirical rule, which associates low-thermal conductivity to the large average atomic radius of the elements composing the alloy.

In the case of the magnets instead a full structural stability analysis, ternary Hull diagrams, was carried out over 36,000 prototypes constructed from elements belonging to the 3d, 4d and 5d periods. This identified about 230 stable compounds, 22 of which baring a finite magnetic moment in the ground state. Then a simple regression based on available experimental data was conducted for estimating the magnetic critical temperature, T_C. The regression was conducted separately for different classes of magnets, yielding a Slater-Pauling curve in the case of Co_2YZ alloys and the empirical Castellitz-Konamata curves for X_2MnZ alloys. The synthesis of four alloys with predicted high-T_C was attempted, resulting in the discovery of Co_2MnTi, a high-moment ferromagnet with a $T_C = 940\,\mathrm{K}$.

Composite materials innovation

Transportation industry has been looking for light-weight materials in their bid to design and build for the environment. Polymer composites are known for their high specific mechanical properties and hence, are now widely used in the aerospace industry. Automobile industry is emulating the success of the aerospace industry. However, the volumes of composites used in the automobile industry are very high compared to the aerospace industry. Hence, for their widespread use in the automobile industry, their environmental credentials should be of significant importance.

Polymer composites are reinforced with either glass fibers or carbon fibers. End-of-use disposal of glass and carbon fiber reinforced composites poses considerable environmental and health challenges. Moreover, the production processes used in making these synthetic fibers are energy intensive. Corona et al. (2016) analyzed the potential of using plant based natural fibers to replace synthetic fibers in composites. Using life cycle assessment (LCA), they concluded that replacement of glass fibers with flax fibers show a general reduction of the environmental impact regardless of the type of application. They also noted that better environmental performance is possible when the combination of fiber and matrix is optimal, and not when the natural fiber content is maximized.

Hervy et al. (2015) conducted LCA of nanocellulose-reinforced advanced fiber composites. They investigated the environmental impact of bacterial cellulose (BC)—and nanofibrillated cellulose (NFC)-reinforced epoxy composites and benchmarked with other viable materials in the same category, such as neat polylactide (PLA) and 30 wt.-% randomly oriented glass fibre-reinforced polypropylene (GF/PP) composites. BC- and NFC-epoxy composites have higher global warming potential (GWP) compared to the neat PLA and GF/PP. However, they found out from the LCA results that the cradle-to-grave GWP of BC- and NFC-composites could be lower than neat PLA when the composites contain over 60% nanocellulose by volume. This suggests that composites with high nanocellulose loading is desirable to produce materials with "greener" credentials. The global market for nanocellulose is estimated at ten billion dollars and the range of applications is growing. The diverse uses include packaging materials, thickening agents, high efficiency filters, automobile components, and mobile electronic devices. LCA provides guidance to the product designers on how best to use nanocellulose-based materials, while improving the sustainability credentials of innovative products.

Nanomaterials innovation

A variety of nanomaterials are already in use in several innovative products. Carbon nanotubes (CNTs) and graphene are

added to plastics. Nano-TiO2 is used in air filter systems. Nano-silver is used in textiles. Nano-WC-Cobalt used in sintered tools. Nano-copper in wood preservatives, nano-SiO2 as food additive, organic pigment in coatings, plastic nanobeads in personal care products, and nanofibers in textiles are some more examples. Bergamaschi et al. (2015) studied the impact and effectiveness of risk mitigation strategies on the insurability of nanomaterial production. They investigated three different nanomaterials, namely ZrO2, TiO2 and multiwalled carbon nanotubes (MWCNT), to obtain an insight into the future applications. They found that by altering the surface chemistry and microstructure, it is possible to modify the hazardous nature of the nanomaterials. In other words, this safety-by-design approach can be emulated in other nanomaterials.

Cerri and Terzi (2016) suggested toolsets to apply information and communication technology (ICT) to improve the life cycle sustainability of manufacturing. It comprises two tools. One is the Life Cycle Optimization Tool that minimizes the life cycle costs and life cycle environmental impacts, according to technical constraints. The tool enables the comparison of different technological solutions. Another is the Life Cycle Data Tool to manage the data collected from the data importing *.csv files acquired via the programmable logic controllers (PLCs) of the different stations or machines using new QLM language. In line with the concepts of Internet of Things (IoT) and Cyber-Physical System (CPS), the proposed toolset are to be positioned between the product lifecycle management (PLM) system and MES (manufacturing execution system) with relatively minimal interoperability constraints. Such ICT-based tools facilitate: (1) real time monitoring of the system; (2) automatic updating of the database thus enabling life cycle cost and environmental evaluations; (3) enabling the comparison between real data and theoretical estimations; and (4) real time insights into the operational behavior (failure rate, availability, reliability, energy consumption, etc.) of the system. These advances are enablers for scientists and engineers to innovate new materials and manufacturing technologies.

Northwestern University researchers, Zhao et al. (2016) described a material genome approach called NanoMine for polymer nanocomposites. The NanoMine infrastructure comprises material database, analysis and simulation tools. It is an open, dynamic and data-driven web-based platform with relational data on composition, microstructure and properties of polymer nanocomposites. Statistical correlations are developed to link processing conditions, quantified microstructure information and macroscopic property response, coupled with image analysis techniques and physics-based simulations. The curated database holds materials properties from the experiments as well as predictions from the physics-based models of polymer nanocomposites. Physical properties, nano- and micro-structures as well as material processing conditions reported in the scientific literature are captured based on curation format and terminology. The data is represented such that it is suitable for further modeling and simulation studies, and used in the materials design for innovative products. NanoMine adopted the Material Data Curator System (MDCS) proposed by NIST for data entry with pre-defined XML schema and customized raw processing, structure and property (p–s–p) parameters so that users can search and retrieve data to conduct their own analysis. It is also enabled with a graphical user interface (GUI). For microstructural analysis, they developed the Niblack Binarization tool which adopts a dynamic local thresholding algorithm to convert input grayscale micrographs into a binary image of separated phases of nanocomposites. Using statistical tool, the binary images are deciphered into quantitative volume fractions and distributions of fibers and fillers. The reconstruction tool and algorithms recreate 3D microstructural images or visualizations from 2D micrographs. They also incorporated Finite Element Analysis (FEA), a physics based continuum model, for the prediction of viscoelastic and dielectric properties of nanocomposites. The commercial software, COMSOL/Abaqus are used for FEA and integrated with database using API and subroutine scripts. It is to be noted that such clear mechanistic modeling may not yet be available for other types of materials, i.e. thermoelectrics, batteries, superconductors, and biomaterials. For such cases, it is reasonable to proceed with phenomenological and empirical models rather than waiting for the development of robust mechanistic models.

NanoMine was built by surveying thirty representative papers on polymer nanocomposites from the literature over the past ten years. NanoMine is developed such that it enabled property prediction for given composition of nanocomposites and material design to obtain a nanocomposite with specified properties. Moreover, as more data from literature becomes available, it is designed to be scalable so as to include more types of nanocomposite materials, such as matrices and reinforcements with different geometry, chemistry and mixing ratios. Moreover, the users can access the Representational State Transfer (REST) API scripts that come with the MDCS. This enables smooth exchange of data between users and the data resource. Further effort is needed to develop forward and inverse models for PSPP relationships (Agrawal and Choudhary 2016).

Hierarchical materials informatics

The significant gap between materials science and design/manufacturing is perhaps most clearly articulated in terms of the vast differences in the length and time scales of the phenomena studied. While design and manufacturing is largely concerned with the predictions of the macroscale (effective) properties and performance of the engineered components

and devices, materials science and engineering is focused on the fundamental understanding of the physical phenomena occurring in the materials multiple length/structure (and time) scales responsible for delivering these properties and performance characteristics. Although the scientific benefits of understanding multiscale materials phenomena have been long appreciated, only recently have we begun to recognize the tremendous potential economic benefits that could accrue from rapid and seamless communication of the materials knowledge into the realm of design/manufacturing.

The afore-mentioned PSPP linkages denote a natural interface for the communications between materials science and design/manufacturing. In many ways, the PSPP linkages can serve a function analogous to the APIs (application program interfaces) used to specify interactions between software components. In other words, a versatile and extensile framework for PSPP linkages can help streamline the communications between materials scientists and designers, and can potentially create the critically needed interoperability between the toolsets used by these two communities. The central challenge in formulating such a framework for PSPP linkages comes from the lack of a rigorous statistical approach for quantifying the hierarchical structure of the material (spanning a multitude of length scales from the atomic to macroscale).

Materials informatics can play a key role in addressing the gap identified above. In particular, a specific branch of materials informatics, called hierarchical materials informatics (Kalidindi 2015), pays particular attention to data-driven (objective) identification of the salient (statistical) measures of the material internal structure, and their use in the establishment of high value, robust, and reliable PSP linkages expressed as low-computational cost reduced-order models. This was accomplished mainly through the use of the formalism of the n-point spatial correlations (Adams et al. 2012) for quantifying the material internal structure, followed by the use of dimensionality reduction approaches (e.g., principal component analyses). Indeed, recent work has demonstrated that a simple workflow template combining these two steps with a suitable framework for formulating reduced-order models (e.g., machine learning) produces highly reliable PSP linkages. Most of the case studies reported can be broadly separated into two groups: (1) communicating salient information from lower length scales to the higher length scales (referred as homogenization), and (2) communicating salient information from higher length scales to lower length scales (referred as localization).

In the homogenization direction, hierarchical materials informatics approaches described above have been successfully employed in relating the complex porous structure of transport layers in Polymer Electrolyte Fuel Cells to their effective diffusivity (Çeçen et al. 2014), the two-phase microstructures in a steel-inclusion material system to their effective plastic properties (Gupta et al. 2015), the development and demonstration of novel high throughput experimental assays for recovering PSP linkages in dual-phase steels (Khosravani et al. 2017), and the extraction of reusable process-structure linkages from highly expensive computational simulations of microstructure evolution in ternary Ag-Cu-Al alloys using phase-field models (Yabansu et al. 2017). In the localization direction, these approaches have been successfully employed in predicting local thermo-elastic stress and strain fields in multiphase composite (Landi et al. 2010; Landi and Kalidindi 2010) and polycrystals (Yabansu et al. 2014 and Yabansu and Kalidindi 2015) local plastic strain rate fields in a two-phase composite (Kalidini et al 2010), and the local compositional fields in spinodal decomposition of metallic alloys (Brough et al. 2016). Although the foundational elements of this emerging new field of hierarchical materials informatics has been laid out in this recent work, much additional work still needs to be done in integrating these elements into an overall material design and development workflow.

Design for environment

Increasingly new products are expected to be designed for the environment with the goal of achieving minimal or zero environmental impact; while meeting the product functions, aesthetics, and costs. Lu et al. (2011) described a process-based sustainable product development (SPD) approach which incorporates life cycle quality (LCQ), life cycle assessment (LCA), and life cycle cost analysis (LCC). A methodology for an integrated life cycle approach to Design for Environment (DfE) was proposed by Low et al. (2014) to ensure that the eco-efficiency of a product over its entire life cycle is considered in the design stage, thereby avoiding the shifting of environmental burdens from one life cycle stage to another. Life cycle evaluation tools such as these allow the estimation of environmental and economic impact of each stage of the product life cycle from cradle to grave, i.e. from raw materials acquisition, through manufacturing, distribution and use, to finally disposal at end-of-life (EoL).

Besides the intrinsic effect on the manufacturing stage (e.g. selection of processes to manufacture product based on materials used), distribution stage (e.g. weight of product manufactured from materials used as a determinant of transportation load) and use stage (e.g. functional efficiency derived from materials used), materials informatics extends to the EoL stage as well. For example, Astrup et al. (2009) studied the environmental performance of plastic wastes of different composition and highlighted its relevance on the selection of EoL treatment, such as recycling, down-cycling or incineration with energy recovery. Recycling is found to be the preferable choice for single component and high

purity plastic, whereas incineration with energy recovery is preferable for mixture owing to the pre-treatment processes required to reach appropriate quality for effective recycling. This illustrates the need for evaluation of environmental performance of a product over its entire life cycle and the role materials informatics plays in the design for environment.

Furthermore, due to the sheer volumes of consumer products and their shorter life cycles, design for environment and the role of materials informatics are being placed in a bigger spotlight. Original equipment manufacturers (OEMs) are on constant search for environmentally benign materials and processes, and potential for reverse manufacturing or closed-loop manufacturing, recycling, upcycling, and zero waste. For OEMs of consumer electronics, in order to manage the thousands of components in each product and often, their suppliers spanning the world, they resort to web-based decision support and evaluation systems. Zhang et al. (2004) reported a web-based system for reverse manufacturing and product environmental impact assessment considering end-of-life disposition of electronic products including desktop computers, laptop and server. The searchable web-based system is developed with Java Servlet and XML (eXtensible Markup Language) which is capable of seamless integration with other systems such as CAD and PDM. Eco-Indicator 99 was used to assess the environmental performance of the product, and the material information is stored in the web-based database. They reported that the system has been tested by a major computer manufacturer. The web-based platform facilitates information sharing by the manufacturers, recyclers and government agencies. It is a multi-tier decision support and evaluation system for operations in remanufacturing and recycling include product disassembly, product recycling, material assessment, environmental impact assessment considering EoL dispositions, product evaluation, and product and material information management. It resulted in reducing environmental impacts by improved materials selection and product design, and focusing on effective recycling processes. Moreover the companies increased profits by optimizing reverse production planning.

Waste-to-resource matching for enablement of industrial symbiosis

Besides the obvious recycling of metals and plastics, other types of wastes can be used as substitutes for raw materials. This can create what is known as industrial symbiosis whereby wastes (or by-products which are traditionally regarded as wastes) are physically exchanged between different companies from within and across industries. To enable industrial symbiosis, the application of materials informatics for waste-to-resource matching could play a significant role (Song et al. 2015; Raabe et al. 2017). Through an ICT-enabled collaboration platform, information and knowledge about whether a waste is recyclable or transformable into a useful resource can be captured and shared among the participants of the industrial symbiosis network. And with the right computational models, algorithms and data, recommendations for what to be exchanged and with whom to establish the exchange can be provided to each participant based on the economic and environmental viability of the exchanges.

However, there are challenges to the widespread use of ICT-enabled systems for enabling industrial symbiosis at present. In a survey by Grant et al. (2010), 17 of such systems were developed during the period 1997–2009. Of these systems, nine are already inactive. The limited success of these systems may be attributed to the challenge in codifying the vast and growing amount of tacit knowledge applied in the determination of waste-to-resource matches. Furthermore, in order to determine a waste-to-resource match, enormous amount of data and information about the multitude of wastes and resources will need to be processed and analyzed—thus, making the task computationally expensive and time-consuming.

Nevertheless, there is one of a few relatively more successful systems that stood out: Core Resource for Industrial Symbiosis Practitioners (CRISP). Developed by the National Industrial Symbiosis Program (NISP) in 2006, it can be considered as one of the most successful, complete and most applied ICT-enabled platform in the domain of industrial symbiosis with a track record of 15,000 projects (companies served) resulting in £1.1 billion savings in total. The aim was to establish a common tool for communication, collaboration and management together in a single place—linking users across all 12 regions in the UK. The main functionalities include the identification of synergy opportunities, the ability to draw on in-house expertise, the visibility of events and activities in different regions and the potential to disseminate best practices. It also includes functionalities beyond the pure opportunity identification and assessment such as relationship management, synergy management, data collection and reporting, communication, collaboration (NISP 2015a). Available information mentions the possibility of interfacing via software integration of Microsoft SharePoint. The data confidentiality charter reveals insights into the nature and use of company data and information (NISP 2015b). Based on these information, waste-to-resource matches are determined using an input-output matching algorithm to search through a taxonomy or classification of wastes and resources. This algorithm seems to be based on mimicking relationships of previous known matches and data gathered from the 15,000 projects carried out all over the globe, which includes countries like Brazil, China, Mexico, South Africa, Canada and across Europe.

Future directions

The above mentioned examples support the notion that the Materials Informatics is an essential bridge between materials science and design/manufacturing. This should not be surprising because data and its transformation into information, knowledge, and wisdom serve as the basic currency for all transactions between cross-disciplinary fields. This recognition gives the emerging field of Materials Informatics a focus and a purpose that is likely to have a transformative impact on the current practices in materials science and engineering as well as design and manufacturing.

More advances can be expected in terms of cheaper sensors, embedded systems, machine learning, cameras and wireless communication interfaces necessary for faster and ubiquitous data capture and ingestion (Kalidindi et al. 2016a, b). Multicomponent nature of the materials data makes it heterogeneous and complex. There is a challenge to aggregate data from multiple data sources into simplified and searchable data, parse unstructured data and manage the nuances of diverse datasets. Moreover, diverse users have different needs. For example, a product designer or a student may just want a single number (e.g. thermal conductivity of a specific material) while an expert in the field may want all the available raw data, specific assumptions and calculations. Hence, the materials informatics approach needs to accommodate heterogeneous data sets and diverse user needs. Ontology-based approach of computer and information sciences can be employed to limit complexity and data curation, thereby formalizing the types, properties and interrelationships of various materials data and information. More importantly, the web-based data infrastructure should be scalable, capable of handling multiple dynamic queries, and flexible programmable access via application programming interfaces (APIs) and learning tools by other users and developers around the world. In other words, material science needs to determine which of the existing database technologies (relational, NoSQL, graph, image, and time series) need to be leveraged.

Materials innovations, facilitated by materials selection based on the life cycle engineering (LCE) and real time materials informatics, innovative product designs, nanotechnology, additive manufacturing, and advanced manufacturing technologies, enable better products with efficient use of materials, and reduce time and cost of materials design and deployment. Reuse, recycling, remanufacturing, cascaded use, redesign, novel waste processing, upcycling and recovery of resources from waste are emerging mitigating options. Circular economy, closed-loop manufacturing and industrial symbiosis are closed-loop perspectives aimed at zero waste of resources. Several factors pose challenges to the desired resource efficiency goals. For example, there are tens of thousands of materials with diverse functional properties, real time performance, and eco-properties, which are often used in combinations and thus, lead to huge varieties of wastes. There is a need to pool, organize, analyze, interpret, and search through vast amounts and diverse sets of real time data on materials and various stages of the product life cycle. For example, the use of this data in the material selection process for new type of batteries, and piezoelectric materials for energy recovery from waste heat. Materials informatics or materials data science helps to bridge the gap between multiscale models and multiscale experiments. Serendipitously inexpensive sensors, faster and cheaper computing power, cloud computing, open source and user friendly apps, big data analytics, fast computational algorithms, data mining, machine learning and artificial intelligence are becoming available. These coupled with deep knowledge of diverse manufacturing processes and materials will enable resource productivity, process efficiency, improved environmental performance of products and perhaps, zero waste. In a nutshell, this is about the materials data-driven opportunities for the future of manufacturing.

Web-based materials informatics platforms are to be developed further such that they are easily used by the globally distributed materials community for accessing and sharing data. Intuitive and visual interfaces are necessary for distributed collaboration to provide visibility and foster trust among stakeholders. Authenticity and traceability of data to the original contributors while beefing up the cyber security and cloud storage capability are necessary. Database platform to cater to the development of new Apps will spur the growth of new businesses.

The materials informatics community needs to work on developing relevant international standards and codes which provide guidance in areas such as materials data exchange, security, design, use or performance of materials, products, processes, services and systems so that they can enhance the reliability of materials data infrastructure and confidence of the users. Materials informatics community also needs to identify and prioritize opportunities and gaps via scholarly engagements.

Conclusions

Materials informatics is gaining traction with the materials scientists, engineers, product designers, innovators, and funding agencies. It is started as web-based searchable repository of materials data. Added functionalities and tools enable selection of materials, design of materials, modeling, simulation and prediction of properties, design for pre-set failure of the product, and evaluation of materials. Further advances in materials informatics assist in driving materials efficiency in manufacturing, matching processes with materials, matching materials with the service and environmental performance,

and designing materials and products for recycling, upcycling and zero waste. Materials efficiency coupled with energy efficiency and water efficiency will contribute to overall resource efficiency in manufacturing, thereby enabling transformation towards sustainable societies. It is hoped that further research and development of materials informatics will lead to greater efficiency in materials use, and uncover fundamental knowledge of the basis of physical, mechanical, electrical, electronic, chemical, biological, and engineering properties. A wider expectation is that the materials informatics will be developed to a point that it is intuitive and easy to use. Moreover, with time, others can even develop apps that makes it accessible to anyone involved in materials and product development. While materials advancements contribute to manufacturing advances and new products; materials informatics will become the materials handbook of modern times.

Acknowledgements Seeram Ramakrishna acknowledges support from Lloyds Register Foundation Grant LRF WBS 265-000-553-597. Surya R. Kalidini acknowledges support from NIST Grant 70NANB14H191. W.C. Lu, Q. Qian and T.Y. Zhang acknowledge support from National Key Research and Development Program of China (2016YFB0700504, and Science and Technology Commission of Shanghai Municipality (Nos. 15DZ2260300 and 16DZ2260600), China. Stefano Sanvito acknowledge support from Science Foundation of Ireland (14/IA/2624 and AMBER Center).

Appendixes

Appendix 1: Mean squared error (MSE) and R^2 statistics

The predictability of a trained machine learning model is usually measured by *RMSE* (*or MSE*) and coefficient of determination (R^2). Their definitions are given by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - e_i)^2}{n}}$$

$$MSE = \frac{\sum_{i=1}^{n}(p_i - e_i)^2}{n}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(e_i - p_i)^2}{\sum_{i=1}^{n}(e_i - \bar{e})^2}$$

where e_i and p_i are respectively the measured and predicted values of a property in interest for test i, n is the number of total tests, and $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i$.

Appendix 2: Voigt–Reuss–Hill (VRH) average

In linear elasticity, stress is linearly proportional to strain and vice versa. This linear relationship is called Hooke's law. For anisotropic materials, Hooke's law takes the forms of $\boldsymbol{\sigma}_{ij} = c_{ijkl}\boldsymbol{\varepsilon}_{kl}$ and $\boldsymbol{\varepsilon}_{ij} = s_{ijkl}\boldsymbol{\sigma}_{kl}$, i, j, k, l = 1, 2, 3, where σ and ε denote the stress and strain tensors, respectively, c is the elastic constant tensor, and s is the elastic compliance tensor. For isotropic materials, there are only two independent elastic constants, although five elastic constants are widely used. The five elastic constants are Young's modulus Y, shear modulus G, bulk modulus K, Poisson ratio ν, and Lame constant λ, and their relations are given by $Y = 2G(1+\nu)$, $K = (3\lambda + 2\nu)/3$, $v = \lambda/[2(G+\lambda)]$. Based on the uniform stress assumption, the Voigt averaged shear modulus and Lame constant are respectively given by $G_{Voigt} = \frac{1}{30}(3c_{ijij} - c_{iijj})$ and $\lambda_{Voigt} = \frac{1}{15}(2c_{iijj} - c_{ijij})$, where the repeated i and j mean the summation over i and j for i, j = 1 2, 3. Then, K_{Voigt} is calculated from G_{Voigt} and λ_{Voigt}. The Voigt averaged elastic moduli are the upper bounds of the elastic moduli. Based on the uniform stress assumption, the Reuss averaged shear modulus and Young's modulus are respectively given by $\frac{1}{G_{Reuss}} = \frac{2}{15}(3s_{ijij} - s_{iijj})$ and $Y_{Reuss} = \frac{1}{15}(2s_{ijij} + s_{iijj})$, where the repeated i and j mean the summation over i and j for i, j = 1 2, 3. Then, K_{Reuss} is calculated from G_{Ruess} and Y_{Reuss}. The Reuss averaged elastic moduli are the lower bounds of the elastic moduli. Voigt–Reuss–Hill average takes the mean of Voigt average and Reuss average, meaning that

$$G_{VRH} = \frac{1}{2}(G_{Voigt} + G_{Reuss}),$$

$$K_{VRH} = \frac{1}{2}(K_{Voigt} + K_{Reuss}).$$

References

Adams, B. L., Kalidindi, S. R., & Fullwood, D. (2012). *Microstructure sensitive design for performance optimization*. Oxford: Butterworth-Heinemann.

Agrawal, A., & Choudhary, A. (2016). Perspective: Materials informatics and big data: Realization of the fourth paradigm of science in materials science. *APL Materials*, *4*(5), 053208. https://doi.org/10.1063/1.4946894.

Agrawal, A., Deshpande, P. D., Cecen, A., Gautham, B. P., Choudhary, A. N., & Kalidindi, S. R. (2014). Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters. *Integrating Materials and Manufacturing Innovation*, *3*, 8. https://doi.org/10.1186/2193-9772-3-8.

Astrup, T., Møller, J., & Fruergaard, T. (2009). Incineration and co-combustion of waste: Accounting of greenhouse gases and global warming contributions. *Waste Management & Research*, *27*(8), 789–799. https://doi.org/10.1177/0734242X09343774.

Balachandran, P. V., Xue, D., Theiler, J., Hogden, J., & Lookman, T. (2016). Adaptive strategies for materials design using uncertainties. *Scientific Reports*, *6* (1966).

Bergamaschi, E., Murphy, F., Poland, C. A., Mullins, M., Costa, A. L., Mcalea, E., et al. (2015). Impact and effectiveness of risk mitigation strategies on the insurability of nanomaterial production: Evidences from industrial case studies. *Wiley Interdisciplinary*

Reviews: Nanomedicine and Nanobiotechnology, *7*(6), 839–855. https://doi.org/10.1002/wnan.1340.

Brough, D. B., Wheeler, D., Warren, J., & Kalidindi, S. R. (2016). Microstructure-based knowledge systems for capturing process-structure evolution linkages. *Current Opinion in Solid State & Materials Science*, *21*, 129–140.

Brownlee, J. (2013) *A tour of machine learning algorithms*. http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/.Accessed 2012.

Carrete, J., Li, W., Mingo, N., Wang, S., & Curtarolo, S. (2014). Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling. *Physical Review X*, *4*, 011019.

Çeçen, A., Fast, T., Kumbur, E. C., & Kalidindi, S. R. (2014). A data-driven approach to establishing microstructure-property relationships in porous transport layers of polymer electrolyte fuel cells. *Journal of Power Sources*, *245*, 144–153.

Cerri, D., & Terzi, S. (2016). Proposal of a toolset for the improvement of industrial systems' lifecycle sustainability through the utilization of ICT technologies. *Computers in Industry*, *81*, 47–54. https://doi.org/10.1016/j.compind.2015.09.003.

Corona, A., Madsen, B., Hauschild, M. Z., & Birkved, M. (2016). Natural fibre selection for composite eco-design. *CIRP Annals Manufacturing Technology*, *65*(1), 13–16. https://doi.org/10.1016/j.cirp.2016.04.032.

de Jong, M., Chen, W., Notestine, R., Persson, K., Ceder, G., Jain, A., et al. (2016). A statistical learning framework for materials science: Application to elastic moduli of k-nary inorganic polycrystalline compounds. *Scientific Reports*, *6*, 34256. https://doi.org/10.1038/srep34256.

de Pablo, J. J., Jones, B., Kovacs, C. L., Ozolins, V., & Ramirez, A. P. (2014). The materials genome initiative, the interplay of experiment, theory and computation. *Current Opinion in Solid State and Materials Science*, *18*(2), 99–117. https://doi.org/10.1016/j.cossms.2014.02.003.

Dean, J. (1990). Lange's handbook of chemistry. *Material and Manufacturing Process*, *5*(4), 687–688.

Dima, A., Bhaskarla, S., Becker, C., Brady, M., Campbell, C., Dessauw, P., et al. (2016). Informatics infrastructure for the materials genome initiative. *JOM*, *68*(8), 2053–2064. https://doi.org/10.1007/s11837-016-2000-4.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journals of American Statistical Association*, *78*, 316–331.

Friedman, J. (2001). Greedy boosting approximation: A gradient boosting machine. *Annals of Statistics*, *29*, 1189–1232. https://doi.org/10.1214/aos/1013203451.

Fischer, C. C., Tibbetts, K. J., Morgan, D., & Ceder, G. (2006). Predicting crystal structure by merging data mining with quantum mechanics. *Nature Materials*, *5*(8), 641–646.

Fürnkranz, J., Gamberger, D., & Lavrač, N. (2012). *Foundations of rule learning*. Berlin: Springer.

Gen, M., & Cheng, R. (1997). *Genetic algorithms and engineering design*. New York: Wiley.

Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C., & Scheffler, M. (2014). Big data of materials science: Critical role of the descriptor. *Physical Review Letters*, *114*(10), 105503.

Grant, G. B., Seager, T. P., Massard, G., & Nies, L. (2010). Information and communication technology for industrial symbiosis. *Journal of Industrial Ecology*, *14*, 740–753.

Gupta, A., Cecen, A., Goyal, S., Singh, A. K., & Kalidindi, S. R. (2015). Structure-property linkages using a data science approach: Application to a non-metallic inclusions/steel composite system. *Acta Materialia*, *91*, 239–254.

Hervy, M., Evangelisti, S., Lettieri, P., & Lee, K.-Y. (2015). Life cycle assessment of nanocellulose-reinforced advanced fibre composites. *Composites Science and Technology*, *118*, 154–162. https://doi.org/10.1016/j.compscitech.2015.08.024.

Jain, A., Persson, K. A., & Ceder, G. (2016). Research Update: The materials genome initiative: Data sharing and the impact of collaborative ab initio databases. *APL Materials*, *4*(5), 053102. https://doi.org/10.1063/1.4944683.

Kalidindi, S. R. (2015). *Hierarchical materials informatics*. Oxford: Butterworth Heinemann.

Kalidindi, S. R., Niezgoda, S. R., Landi, G., Vachhani, S., & Fast, T. (2010). A novel framework for building materials knowledge systems. *Computers, Materials & Continua*, *17*, 103–125.

Kalidindi, S. R., Brough, D. B., Li, S., Cecen, A., Blekh, A. L., Congo, F. Y. P., et al. (2016a). Role of materials data science and informatics in accelerated materials innovation. *MRS Bulletin*, *41*(8), 596–602. https://doi.org/10.1557/mrs.2016.164.

Kalidindi, S. R., & De Graef, M. (2015). Materials data science: Current status and future outlook, annual review of materials research. *Annual Reviews*, *45*(1), 171–193. https://doi.org/10.1146/annurev-matsci-070214-020844.

Kalidindi, S. R., Medford, A. J., & McDowell, D. L. (2016b). Vision for data and informatics in the future materials innovation ecosystem. *JOM*, *68*, 2126–2137.

Kaufman, L., & Ågren, J. (2014). CALPHAD, first and second generation—Birth of the materials genome. *Scripta Materialia*, *70*, 3–6. https://doi.org/10.1016/j.scriptamat.2012.12.003.

Khosravani, A., Cecen, A., & Kalidindi, S. R. (2017). Development of high throughput assays for establishing process-structure-property linkages in multiphase polycrystalline metals: Application to dual-phase steels. *Acta Materialia*, *123*, 55–69.

Landi, G., Niezgoda, S. R., & Kalidindi, S. R. (2010A). Multi-scale modeling of elastic response of three-dimensional voxel-based microstructure datasets using novel DFT-based knowledge systems. *Acta Materialia*, *58*, 2716–2725.

Landi, G., & Kalidindi, S. R. (2010B). Thermo-elastic localization relationships for multi-phase composites. *Computers, Materials & Continua*, *16*, 273–293.

Le, T. C., & Winkler, D. A. (2016). Discovery and optimization of materials using evolutionary approaches. *Chemical Reviews*, *116*(10), 6107–6132. https://doi.org/10.1021/acs.chemrev.5b00691.

Lookman, T., Balachandran, P. V., Xue, D., Hogden, J., & Theiler, J. (2016). Statistical inference and adaptive design for materials discovery. *Current Opinion in Solid State and Materials Science*,. https://doi.org/10.1016/j.cossms.2016.10.002.

Low, J. S. C., Lu, W. F., & Song, B. (2014). Methodology for an integrated life cycle approach to design for environment. *Key Engineering Materials*,. https://doi.org/10.4028/www.scientific.net/KEM.572.20.

Lu, B., Zhang, J., Xue, D., & Gu, P. (2011). Systematic lifecycle design for sustainable product development. *Concurrent Engineering*,. https://doi.org/10.1177/1063293X11424513.

McDowell, D. L., & Kalidindi, S. R. (2016). The materials innovation ecosystem: A key enabler for the materials genome initiative. *MRS Bulletin*, *41*(4), 326–337. https://doi.org/10.1557/mrs.2016.61.

Mitchell, T. M. (1997). *Machine Learning*. New York City: McGraw Hill.

Mulholland, G. J., & Paradiso, S. P. (2016). Perspective: Materials informatics across the product lifecycle—Selection, manufacturing, and certification. *APL Materials*, *4*(5), 053207. https://doi.org/10.1063/1.4945422.

National Science and Technology Council. (2011). *Materials genome initiative for global competitiveness*. Washington: National Science and Technology Council.

NISP. (2015a). *A brief introduction to CRISP*. National Industrial Symbiosis Programme. http://sdrn.policystudiesinstitute.org.uk/sites/default/files/events/Paul Innes CRI (Accessed October 25, 2015).

NISP. (2015b). *Confidentiality Charter CRISP*. National Industrial Symbiosis Programme. Available at: https://www.tees.ac.uk/docs/DocRepo/Clemance/NISPConfidentialtyCharter.pdf (Accessed: 26 October 2015).

NIST. (2013). *Materials informatics*. National Institute of Standards and Technology. https://www.nist.gov/programs-projects/materials-informatics. Accessed October 30, 2016.

Panchal, J. H., Kalidindi, S. R., & McDowell, D. L. (2013). Key computational modeling issues in integrated computational materials engineering. *Journal of Computer-Aided Design*, *45*, 4–25.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *27*(8), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159.

Phillips, C. L., & Littlewood, P. (2016). Preface: Special topic on materials genome. *APL Materials*, *4*(5), 2014–2016. https://doi.org/10.1063/1.4952608.

Puchala, B., Tarcea, G., Marquis, E. A., Hedstrom, M., Jagadish, H. V., & Allison, J. E. (2016). The materials commons: A collaboration platform and information repository for the global materials community. *JOM*, *68*(8), 2035–2044. https://doi.org/10.1007/s11837-016-1998-7.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Raabe, B., Low, J. S. C., Juraschek, M., Herrmann, C., Tjandra, T. B., Ng, Y. T., et al. (2017). Collaboration platform for enabling industrial symbiosis? Application of the by-product exchange network model. *Procedia CIRP*, *61*, 263–268. https://doi.org/10.1016/j.procir.2016.11.225.

Raccuglia, P., Elbert, K. C., Adler, P. D. F., Falk, C., Wenny, M. B., Mollo, A., et al. (2016). Machine-learning-assisted materials discovery using failed experiments. *Nature*, *533*(7601), 73–76. https://doi.org/10.1038/nature17439.

Rajan, Jose, & Seeram, Ramakrishna, (2018). Materials 4.0: Materials big data enabled materials discovery. *Applied Materials Today*. https://doi.org/10.1016/j.apmt.2017.12.015

Rajan, K. (2015). Materials informatics: The materials gene and big data. *Annual Review of Materials Research*, *45*(1), 153–169. https://doi.org/10.1146/annurev-matsci-070214-021132.

Rodgers, J. R., & Cebon, D. (2006). Materials informatics. *MRS Bulletin,31*(12), 975–980. https://www.cambridge.org/core/article/materials-informatics/4DDA16B3B93C616EBAE618445488A09B.

Samuel, A. L. (1967). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *11*(6), 601–617.

Sanvito, S., Oses, C., Xue, J., Tiwari, A., Zic, M., Archer, T., et al. (2017). Accelerated discovery of new magnets in the Heusler alloy family. *Science Advances*, *3*, e1602241.

Shalev-Shwartz, S. (2011). Online learning and online convex optimization. *Foundations and Trends in Machine Learning.*, *4*(2), 107–194. https://doi.org/10.1561/2200000018.

Song, B., Yeo, Z., Low, J. S. C., Koh, D. J., Kurle, D., Cerdas, F., et al. (2015). A big data analytics approach to develop industrial symbioses in large cities. *Procedia CIRP*, *29*, 450–455. https://doi.org/10.1016/j.procir.2015.01.066.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological),58*(1), 267–88. http://www.jstor.org/stable/2346178.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Wang, Z., et al. (2014). Integrated materials design and informatics platform within the materials genome framework. *Chinese Science Bulletin*, *59*(15), 1755–1764. https://doi.org/10.1007/s11434-014-0225-6.

Xiong, P., Ji, X., Zhao, X., Lv, W., Liu, T., & Lu, W. (2015). Materials design and control synthesis of the layered double hydroxide with the desired basal spacing. *Chemometrics and Intelligent Laboratory Systems*, *144*, 11–16. https://doi.org/10.1016/j.chemolab.2015.03.005.

Xue, D., Balachandran, P. V., Hogden, J., Theiler, J., Xue, D., & Lookman, T. (2016). Accelerated search for materials with targeted properties by adaptive design. *Nature Communications*, *7*, 11241.

Xue, D., Xue, D., Yuan, R., Zhou, Y., Balachandran, P. V., Ding, X., et al. (2017). An informatics approach to transformation temperatures of NiTi-based shape memory alloys. *Acta Materialia*, *125*, 532–541. https://doi.org/10.1016/j.actamat.2016.12.009.

Yabansu, Y. C., & Kalidindi, S. R. (2015). Representation and calibration of elastic localization kernels for a broad class of cubic polycrystals. *Acta Materialia*, *94*, 26–35.

Yabansu, Y. C., Patel, D. K., & Kalidindi, S. R. (2014). Calibrated localization relationships for elastic response of polycrystalline aggregates. *Acta Materialia*, *81*, 151–160.

Yabansu, Y. C., Steinmetz, P., Hotzer, J., Kalidindi, S. R., & Nestler, B. (2017). Extraction of reduced-order process-structure linkages from phase-field simulations. *Acta Materialia*, *124*, 182–194.

Zhang, H. C., Li, J., Shrivastava, P., Whitley, A., & Merchant, M. E. (2004). A web-based system for reverse manufacturing and product environmental impact assessment considering enf-of-life dispositions. *CIRP Annals Manufacturing Technology*, *53*(1), 5–8. https://doi.org/10.1016/S0007-8506(07)60632-5.

Zhao, Y. H., Abraham, M. H., & Zissimos, A. M. (2003). Determination of McGowan volumes for ions and correlation with van der Waals volumes. *Journal of Chemical Information & Computer Sciences*, *43*(6), 1848–1854.

Zhao, H., Li, X., Zhang, Y., Schadler, L. S., Chen, W., & Brinson, L. C. (2016). Perspective: NanoMine: A material genome approach for polymer nanocomposites analysis and design. *APL Materials*, *4*(5), 053204. https://doi.org/10.1063/1.4943679.