

Challenges and Advances
in Computational Chemistry and Physics 24
Series Editor: Jerzy Leszczynski

Kunal Roy *Editor*

Advances in QSAR Modeling

Applications in Pharmaceutical,
Chemical, Food, Agricultural and
Environmental Sciences

 Springer

Challenges and Advances in Computational Chemistry and Physics

Volume 24

Series editor

Jerzy Leszczynski
Department of Chemistry and Biochemistry
Jackson State University, Jackson, MS, USA

This book series provides reviews on the most recent developments in computational chemistry and physics. It covers both the method developments and their applications. Each volume consists of chapters devoted to the one research area. The series highlights the most notable advances in applications of the computational methods. The volumes include nanotechnology, material sciences, molecular biology, structures and bonding in molecular complexes, and atmospheric chemistry. The authors are recruited from among the most prominent researchers in their research areas. As computational chemistry and physics is one of the most rapidly advancing scientific areas such timely overviews are desired by chemists, physicists, molecular biologists and material scientists. The books are intended for graduate students and researchers.

More information about this series at <http://www.springer.com/series/6918>

Kunal Roy
Editor

Advances in QSAR Modeling

Applications in Pharmaceutical, Chemical,
Food, Agricultural and Environmental
Sciences

 Springer

Editor
Kunal Roy
Department of Pharmaceutical Technology
Jadavpur University
Kolkata
India

ISSN 2542-4491 ISSN 2542-4483 (electronic)
Challenges and Advances in Computational Chemistry and Physics
ISBN 978-3-319-56849-2 ISBN 978-3-319-56850-8 (eBook)
DOI 10.1007/978-3-319-56850-8

Library of Congress Control Number: 2017937135

© Springer International Publishing AG 2017

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Quantitative structure–activity/property relationship (QSAR/QSPR) models correlate biological activity (therapeutic or toxic) or other properties of chemicals/pharmaceuticals/toxicants/environmental pollutants with molecular structure information using chemometric and cheminformatic tools. These models have relevant applications in diverse disciplines like materials property modeling, environmental fate modeling, risk assessment, drug design, ADME/T modeling, food chemical and agrochemical design, nanomaterials design, etc. The research output in the mentioned areas is of paramount importance for data gap filling in case of non-availability of experimental data. Thus, model derived predictions have applications for regulatory purposes in chemical industries and for new analogue design in pharmaceutical, food and agricultural industries.

QSAR models aim to explore the specific quantitative relationships between a target activity (or property) and a set of structural and physicochemical features encoded within some numerical quantities popularly known as descriptors. Such models are helpful in predicting the target response for new molecules falling within the applicability domain of the developed models. It is also possible to derive a mechanistic interpretation of the structure–activity relationships, especially from models which have been developed using descriptors with definite physicochemical meaning. The objective of structure–activity/property modeling is to analyze and detect the determining factors for the measured activity/property for a particular biological and/or chemical system in order to have an insight of the mechanism and behavior of the studied system. It is now very important to validate the developed QSAR models in order to check reliability of predictions for new compounds. The Organization for Economic Co-operation and Development (OECD) has recommended a set of five point guidelines for QSAR/QSPR model development and validation, especially for regulatory purposes.

QSAR has long been used in medicinal chemistry for lead optimization and drug design. QSAR increases the probability of success in finding an optimum lead with desired pharmacokinetic profile thus avoiding tedious experiments with thousands of compounds with less potential to become successful and hence avoiding colossal expenditure. QSAR is also a very popular tool for risk assessment of chemicals in

the absence of experimental data. Such approach is used by the United States Environmental Protection Agency and also encouraged in the European Union's REACH legislation. QSAR techniques are in consonance with the '3Rs concept' related to the moral principle regarding the use of sentient animals. As a result of increasing chemicals uses, applications of analogues, SAR and QSAR approaches by global Governmental organizations have increased.

There are four main areas where QSARs may be applied by governmental regulatory agencies:

- (i) Prioritization of existing chemicals for further testing or assessment
- (ii) Classification and labelling of new chemicals
- (iii) Risk assessment of new and existing chemicals
- (iv) Filling of data gaps.

In addition to the applications in chemical risk assessment and pharmaceutical development, the QSPR/QSAR findings can also be used to screen compounds with specific applications such as food additives, antioxidants and nanomaterials.

This volume aims at describing the fundamentals of QSAR modeling and showcasing some recent advancements of QSAR applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences. There are 15 chapters on different topics of QSAR theory, methods and applications included in this book.

Chapter "[Towards the Revival of Interpretable QSAR Models](#)" authored by Chanin Nantasenamat and others gives a good introduction of the key topic QSAR. It highlights the basic steps of model development and validation, discusses various molecular descriptors and statistical techniques for model development. This chapter also discusses key issues influencing and contributing to the interpretability of QSAR models.

Chapter "[The Use of Topological Indices in QSAR and QSPR Modeling](#)" authored by John C. Dearden gives an overview of topological descriptors and their use in QSAR/QSPR studies. The author also mentions about biodescriptors, chirality and software availability in connection with topological descriptors.

Chapter "[Which Performance Parameters Are Best Suited to Assess the Predictive Ability of Models?](#)" has been authored by Karoly Heberger and co-authors. The authors have revisited the debate topic of choice of external validation versus cross-validation as a better tool for judging the predictive potential of QSAR/QSPR models. Using the sum of ranking differences (SRD) methodology coupled with ANOVA, the authors claim the superiority of cross-validation, at least in the case studies reported by them.

Chapter "[Structural, Physicochemical and Stereochemical Interpretation of QSAR Models Based on Simplex Representation of Molecular Structure](#)" authored by Victor Kuzmin and others shows the applicability of simplex descriptors to development of interpretable models using several case studies. The authors also demonstrate the applicability of the SiRMS approach to stereochemical interpretation.

Chapter "[The Maximum Common Substructure \(MCS\) Search as a New Tool for SAR and QSAR](#)" is authored by Giuseppina Gini and others. This chapter discusses relevance of a similarity measure exclusively based on the maximum

common substructure and its implementation in a new software tool developed by the authors and integrated into the ToxRead software. The authors' approach can be used for read across, where only local information about one or two similar molecules is used, or in assessing the prediction of QSAR results, or in refining the results of SAR systems that apply structural alerts.

Chapter “[Generative Topographic Mapping Approach to Chemical Space Analysis](#)” is written by Alexandre Varnek and others. This chapter describes Generative Topographic Mapping (GTM) and its application as a predictive tool for analysis of chemical space. The strengths of GTMs in chemical space navigation and analysis are critically reviewed.

Chapter “[On Applications of QSARs in Food and Agricultural Sciences: History and Critical Review of Recent Developments](#)” authored by Jerzy Leszczynski and others) presents the currently available information on diverse groups of molecules with applications in agriculture and food science that have been subjected to *in silico* modeling studies. The authors have also enlisted available agrochemical, food and flavor databases along with an extensive list of software tools and online resources for QSAR and other related *in silico* modeling studies.

Chapter “[Quantitative Structure-Epigenetic Activity Relationships](#)” authored by Jose Medina-Franco and others) has analyzed the progress of QSAR models developed for compound databases screened with epigenetic targets. This chapter also analyzes epigenetic activity landscape modeling, activity cliffs, and activity cliff generators and their relevance to develop QSAR models.

Chapter “[QSAR/QSPR Modeling in the Design of Drug Candidates with Balanced Pharmacodynamic and Pharmacokinetic Properties](#)” is authored by Anna Tsantili-Kakoulidou and others. It showcases the application of QSAR/QSPR in drug discovery process. This chapter discusses several case studies related to application of QSAR in modeling pharmacodynamics and pharmacokinetics of drug substances.

Chapter “[Strategy for Identification of Nanomaterials' Critical Properties Linked to Biological Impacts: Interlinking of Experimental and Computational Approaches](#)” authored by Iseult Lynch and others discusses on physicochemical properties of nanomaterials in connection with their toxicological outcome and application of QSAR in prediction of nanomaterial uptake and toxicity. This chapter also highlights the gaps between measured physicochemical parameters and calculated QSAR descriptors for nanomaterials.

Chapter “[In Silico Approaches for the Prediction of In Vivo Biotransformation Rates](#)” is authored by Ester Papa and others. This chapter illustrates the development and application of *in silico* models for *in vivo* biotransformation rates and half lives of chemicals. This chapter also describes the complementary role of *in vitro* biotransformation rate estimation and the subsequent *in vitro*-to-*in vivo* extrapolation calculations for refining bioaccumulation model predictions.

Chapter “[Development of Monte Carlo Approaches in Support of Environmental Research](#)” written by Emilio Benfenati and others shows application of the CORAL software for evaluation of environmental effects of various chemical compounds. The mechanistic interpretation and domain of applicability of the models for

various environmentally important endpoints was also discussed from different case studies.

Chapter “[Environmental Toxicity of Pesticides, and Its Modeling by QSAR Approaches](#)” is authored by A. Amrane and others. The chapter reviews pollution by pesticides and their effects on the entire ecosystem. A critical review of QSAR models for the prediction of the toxicity of pesticides is also provided.

Chapter “[Counter-Propagation Artificial Neural Network Models for Prediction of Carcinogenicity of Non-congeneric Chemicals for Regulatory Uses](#)” written by N. Fjodorova and others focuses on QSAR models for prediction of carcinogenic potency based on counter-propagation artificial neural network algorithm. These models were developed in the scope of CAESAR and PROSIL projects and implemented in online available internet platform VEGA.

Chapter “[Big Data in Structure-Property Studies—From Definitions to Models](#)” authored by Jaroslaw Polanski discusses what big data is and how important big data can be in drug design. This chapter also analyzes the big data types that are available in drug design as well as the methods that are used for their analyses.

I hope that this collection of 15 chapters will be helpful to the researchers working in the field of QSAR modeling. I am especially thankful to the Series editor Prof. Jerzy Leszczynski for his help during development of this book and to the publisher for bringing out this volume.

Kolkata, India

Kunal Roy

Contents

Part I Theory

Towards the Revival of Interpretable QSAR Models	3
Watshara Shoombuatong, Philip Prathipati, Wiwat Owasirikul, Apilak Worachartcheewan, Saw Simeon, Nuttapat Anuwongcharoen, Jarl E.S. Wikberg and Chanin Nantasenamat	
The Use of Topological Indices in QSAR and QSPR Modeling	57
John C. Dearden	
Which Performance Parameters Are Best Suited to Assess the Predictive Ability of Models?	89
Károly Héberger, Anita Rácz and Dávid Bajusz	

Part II Methods

Structural, Physicochemical and Stereochemical Interpretation of QSAR Models Based on Simplex Representation of Molecular Structure	107
P. Polishchuk, E. Mokshyna, A. Kosinskaya, A. Muats, M. Kulinsky, O. Tinkov, L. Ognichenko, T. Khristova, A. Artemenko and V. Kuz'min	
The Maximum Common Substructure (MCS) Search as a New Tool for SAR and QSAR	149
Azadi Golbamaki, Alessio Mauro Franchi and Giuseppina Gini	
Generative Topographic Mapping Approach to Chemical Space Analysis	167
Dragos Horvath, Gilles Marcou and Alexandre Varnek	

Part III Applications

On Applications of QSARs in Food and Agricultural Sciences: History and Critical Review of Recent Developments	203
Supratik Kar, Kunal Roy and Jerzy Leszczynski	
Quantitative Structure-Epigenetic Activity Relationships	303
Mario Omar García-Sánchez, Maykel Cruz-Monteagudo and José L. Medina-Franco	
QSAR/QSPR Modeling in the Design of Drug Candidates with Balanced Pharmacodynamic and Pharmacokinetic Properties	339
George Lambrinidis, Fotios Tsopelas, Costas Giaginis and Anna Tsantili-Kakoulidou	
Strategy for Identification of Nanomaterials' Critical Properties Linked to Biological Impacts: Interlinking of Experimental and Computational Approaches	385
Iseult Lynch, Antreas Afantitis, Georgios Leonis, Georgia Melagraki and Eugenia Valsami-Jones	
In Silico Approaches for the Prediction of In Vivo Biotransformation Rates	425
Ester Papa, Jon A. Arnot, Alessandro Sangion and Paola Gramatica	
Development of Monte Carlo Approaches in Support of Environmental Research	453
Alla P. Toropova, Andrey A. Toropov, Emilio Benfenati, Robert Rallo, Danuta Leszczynska and Jerzy Leszczynski	
Environmental Toxicity of Pesticides, and Its Modeling by QSAR Approaches	471
Mabrouk Hamadache, Abdeltif Amrane, Othmane Benkortbi, Salah Hanini, Latifa Khaouane and Cherif Si Moussa	
Counter-Propagation Artificial Neural Network Models for Prediction of Carcinogenicity of Non-congeneric Chemicals for Regulatory Uses	503
N. Fjodorova, M. Novic, S. Zuperl and K. Venko	
Big Data in Structure-Property Studies—From Definitions to Models	529
Jaroslaw Polanski	
Index	553

Part I

Theory

Towards the Revival of Interpretable QSAR Models

Watshara Shoombuatong, Philip Prathipati, Wiwat Owasirikul,
Apilak Worachartcheewan, Saw Simeon, Nuttapat Anuwongcharoen,
Jarl E. S. Wikberg and Chanin Nantasenamat

Abstract Quantitative structure-activity relationship (QSAR) has been instrumental in aiding medicinal chemists and physical scientists in understanding how modification of substituents at different positions on a molecular structure exert its influence on the observed biological activity and physicochemical property, respectively. QSAR has received great attention owing to its predictive capability and as such efforts had been directed toward obtaining models with high prediction performance. However, to be useful QSAR models need to be informative and interpretable in which the underlying molecular features that contribute to the increase or decrease of the biological activity are revealed by the model. Thus, the aim of this chapter is to briefly review the general concepts of QSAR modeling, its development and discussions on key issues influencing and contributing to the interpretability of QSAR models.

W. Shoombuatong and P. Prathipati
These authors contributed equally to this work.

W. Shoombuatong · S. Simeon · N. Anuwongcharoen · C. Nantasenamat (✉)
Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology,
Mahidol University, Bangkok 10700, Thailand
e-mail: chanin.nan@mahidol.edu

P. Prathipati
National Institutes of Biomedical Innovation, Health and Nutrition,
Osaka 567-0085, Japan

W. Owasirikul
Department of Radiological Technology, Faculty of Medical Technology,
Mahidol University, Bangkok 10700, Thailand

A. Worachartcheewan
Department of Community Medical Technology, Faculty of Medical Technology,
Mahidol University, Bangkok 10700, Thailand

J.E.S. Wikberg
Department of Pharmaceutical Biosciences, BMC, Uppsala University,
SE-751 24 Uppsala, Sweden

Keywords Quantitative structure-activity relationship · Quantitative structure-property relationship · Proteochemometrics · Data mining · Machine learning · Cheminformatics · Chemogenomics · QSAR · QSPR · Interpretable · Drug discovery · Drug design

1 Introduction

Quantitative structure-activity relationship (QSAR) can be considered to be one of the pillars for driving drug discovery efforts forward by enabling practitioners to make sense of the big data from bioactivity assays of chemical library (Nantasenamat et al. 2009, 2010; Cherkasov et al. 2014). Computer-aided drug design or simply computational drug design is essentially comprised of four major levels: (i) fragment, (ii) ligand, (iii) structure and (iv) systems based approaches (Nantasenamat and Prachayasittikul 2015). QSAR is a ligand-based approach meaning that it primarily makes use of information derived from ligands that does not require the need for details of the target protein. Thus, ligand-based approaches are particularly suited in situations where there is negligible information on the biological target. The reasons for using QSAR and quantitative structure-property relationship (QSPR) models are many: (i) to reduce time and cost; (ii) to rationally predict biological, pharmaceutical, physical and chemical activities/properties; (iii) to aid experimental scientists by providing the collective wisdom learned from previous big data; (vi) to shed light on the mechanism of action for biological activities of interest. QSAR/QSPR has found wide applications in the life sciences (Prachayasittikul et al. 2015) (e.g. biology, agriculture and medicine) as well as the physical sciences (Katritzky et al. 2010) (e.g. organic chemistry, physical chemistry, materials sciences). In drug discovery, QSAR has been successfully applied in the prediction of $\log P$ and pK_a values as well as absorption, distribution, metabolism, excretion and toxicity (ADMET) properties (Khan and Sylte 2007). It is indeed a difficult task to design a drug that exert activity toward the target protein(s) of interest while at the same time show proper uptake, metabolism, excretion and be devoid of toxicity. To aid medicinal chemists in understanding the origin of ADMET properties Gleeson proposed a set of simple and interpretable rules through the use of principal component analysis of simple descriptors (e.g. molecular weight, $\log P$, ionization state, etc.) (Gleeson 2008).

The robustness of QSAR relies on its capability to predict the biological activities or chemical properties of interests by learning from retrospective experimental data sets. Particularly, each compound in a chemical library is quantitatively or qualitatively described by a set of molecular descriptors and such vector of descriptors (also known as independent variables in statistics) are mathematically correlated with the biological or chemical endpoint of interest (i.e. pIC_{50} , $\log P$, etc.) via traditional multivariate analysis or machine learning algorithm. However, it is worthy to note that QSAR models is only as good as the data that was used to train it and in spite of its predictive capability it should not be viewed as a replacement of domain knowledge

of scientists but rather should be considered as a complementary tool for aiding the decision-making process.

In spite of its widespread usage, it seems that the full potential for QSAR models has not yet been achieved as current efforts are localized on generating models with good predictive performance at the cost of vague or uninterpretable models. Most robust machine learning algorithms are so-called *black box* since the underlying features contributing to the variation in the endpoint values are not accessible to practitioners. To be of benefit for the experimental biologist or chemist, models need to be transparent such that the underlying important features are revealed. Moreover, features describing the general or unique characteristics of compounds needs to be unambiguous, interpretable and easily comprehensible. Upstream to the issue of interpretability is the accessibility or the know-how on the development of robust QSAR models. Nowadays, the construction of QSAR models may seem to be a trivial and mainstream task in computational drug design. However, a robust, reliable and reproducible model can only be achieved through careful data curation and analysis, which certainly requires the expertise of trained practitioners. This is particularly true as not all starting data set is *modelable* or may not always yield promising results right out of the box owing to several inherent issues that will be discussed in this chapter.

2 Brief History of QSAR

More than a century ago, QSAR was developed by several research groups. The precursor to the birth of QSAR began in 1863 when Cros (Cros 1863) observed that there exists an inverse correlation between toxicity and water solubility. Particularly, the toxicity of alcohols toward mammals increased as the water solubility of alcohols decreased. Shortly after, Crum-Brown and Fraser (1868) reported that there was a correlation between chemical substituents and their physiological properties. Later in the 1890s, Hans Horst Meyer reported that the toxicity of organic compounds depended on their lipophilicity (Borman 1990; Lipnick 1991). Subsequently, the linear correlation between lipophilicity (e.g. oil-water partition coefficients) and biological properties was investigated. Louis Hammett (Hansch et al. 1991) investigated the relationship between electronic properties of organic acids and bases with their equilibrium constants and reactivity. These early studies form the basis for the development of modern QSAR by establishing the idea that molecular structures directly influenced the endpoint (i.e. biological activity and chemical property) of interest. In 1962, Hansch et al. (1962) formally coined the term QSAR and laid its initial foundations by investigating the structure-activity relationship (SAR) of plant growth regulators and pesticides and their dependency on Hammett constants (Hammett 1937) and hydrophobicity (Gallup et al. 1952).

The Free-Wilson model (Free and Wilson 1964) is a simple and efficient method for the quantitative description of SAR. It explains the variation in a series of congeneric compounds using the presence or absence of substituents or functional

groups as molecular descriptors. It is the only numerical method that directly relates structural features with biological properties, which is in contrast to Hansch analysis where physicochemical properties are correlated with biological activity values (Kubinyi 1988). Nevertheless, both approaches are closely interrelated, not only from a theoretical point of view but also in their practical applicability (Kubinyi 1988). In many cases both models were combined to afford a mixed approach that includes Free-Wilson type parameters for describing the activity contributions of certain structural modifications and physicochemical parameters for describing the effect of substituents on the biological activity (Kubinyi 1988; Wei et al. 2001). Many successful applications, especially from the work of Hansch and his group (Verma and Hansch 2009; Hansch et al. 2002; Kurup et al. 2000; Gao et al. 1999; Selassie et al. 2002; Kurup et al. 2001; Hansch and Gao 1997; Kurup et al. 2001; Hansch et al. 1996; Hadjipavlou-Litina et al. 2004; Garg et al. 1999, 2003) on the SAR of enzyme inhibitors, demonstrated that this combined model affords stellar performance for classical QSAR (Hansch 2011). Several variations to Free-Wilson approach have been developed and recently found useful applications in fragment-based drug design (Eriksson et al. 2014; Chen et al. 2013; Radoux et al. 2016).

The field of QSAR modeling had evolved progressively and this encompasses two radical transformations as follows:

1. Paradigm shift from the *classical* to the *non-classical* QSAR approach (Fujita and Winkler 2016). The former is based on a small set of congeneric series of compounds that usually have a single mode of action while the latter is based on large, heterogeneous and non-congeneric data set that may contain several mode of actions.
2. Paradigm shift of QSAR models (Nantasenamat et al. 2009, 2010; Cherkasov et al. 2014) that considers the SAR of *several compounds against a single target protein* to the so-called proteochemometric model (Cortes-Ciriano et al. 2015; Qiu et al. 2016) (sometimes referred to as computational chemogenomics) that investigates the SAR of *several compounds against several target proteins*.

3 How Far Can QSAR Take Us: Can It Really Bring a Drug to Market?

QSAR modeling have evolved from concept to initial hype followed by skepticism thereby leading to the identification of their pitfalls and caveats to a moderation of their expectations (Doweyko 2008). QSAR models are routinely used in the prediction of physicochemical properties (e.g. $\log P$, pK_a and solubility) as well as pharmacokinetic and toxicity endpoints (e.g. permeability, plasma protein binding, liver toxicity, carcinogenicity, seizure and off-target activities). However, their usage for actual lead identification and optimization phase has remained quite limited. The skepticism from medicinal chemists towards QSAR models stems from the inability of descriptor based QSAR models (constructed using fingerprints and various

topological descriptors) to rationalize activities in terms of simple, meaningful and constructive ways that can clearly provide details on what modifications should be made to the chemical structure that can afford activity enhancement. Furthermore, with better ability to assimilate data from human readable patents and publications of SAR data in concomitant with better understanding of the isosteric concept, medicinal chemists are better able to capture the underlying principles of SAR and make synthetically feasible and conservative predictions. However, many encouraging signs are beginning to appear as more robust machine learning algorithm and interpretable molecular descriptors are being developed. It is still early to predict the potential of QSAR modeling for bringing a drug to market since they are used in the early stages of a drug discovery project. With the ever increases in the availability of clinical and adverse effect data, the use of QSAR modeling together with complementary computational approaches (e.g. cheminformatics, computational chemistry, molecular docking, molecular dynamics, etc.) helps improve the odds of bringing a drug to market. QSAR modeling in combination with other computer-aided drug design techniques have already shown numerous success stories as summarized in an excellent report by Kubinyi (2006).

3.1 Why Does QSAR Fail?

QSAR modeling, like many other research disciplines, has had its fair share of ups and downs. Many predicted the eventual demise of QSAR due to the advances in synthetic chemistry techniques (e.g. combinatorial chemistry) and assay attributes (e.g. automation and miniaturization). Drug discovery researchers dissolution with QSARs is rooted in the fact that it has yet to demonstrate a robust ability to predict the desired biological activities. The disappointing results from QSAR models in certain situation can be attributed to features obtained by chance correlation, rough response surfaces, incorrect functional forms and overtraining (Johnson 2008; Doweiko 2008). Particularly, rough response surfaces are an inherent characteristic of SAR data sets that nevertheless significantly affect the QSAR model predictions. For instance, most aminergic GPCR ligands' agonistic activities correlate with their pK_a and in many instances an order of magnitude change in the pK_a results in a comparable or even a multi-fold change in the biological activity. Such conservative change in the chemical structure leading to a large change in the activity are often not captured by QSAR models which rely heavily on statistical approaches to capture the features that cause the biological responses. On the other hand, a chemist quickly grasps the trend using rational thought, controlled experiments and personal observation assisted by prior knowledge of the protein's structure-function relationships. This over-reliance on statistical procedures by QSAR researchers for feature selection and data modeling has led to the identification of features that may have no mechanistic role in modulating the activities but might have correlated by chance. The excessive emphasis on machine learning has also resulted in model overfitting, models that uses the incorrect functional forms and/or highly predictive

models with vague or little interpretability. Hence, the resulting QSAR models do not reflect the reality of the binding or modulation event, which causes the predictions to eventually fail. Thus, to derive meaningful hypothesis, practitioners should not blindly rely on results from computational models but should view the results as hints or guides for supporting their own decision-making process (Nantasenamat and Prachayasittikul 2015). Thus, it is recommended to implement some form of expert knowledge guided component in the QSAR workflows such that new solutions are built upon prior knowledge of targets and their modulation (Saxena and Prathipati 2003). In fact, such data-driven approach as implemented in the HADDOCK docking software (Vries et al. 2010) relies on prior biochemical and biophysical data to drive the docking simulations. Moreover, several recent blinded genomic challenges for phenotype prediction such as sbvImprover (Tarca et al. 2013) and DREAM (Costello et al. 2014) also suggests that the inclusion of prior knowledge can significantly enhance the predictive power while consuming minimal computational resources. In this context, the use of interpretable molecular descriptors aided by transparent machine learning models can greatly alleviate the existing problems of QSAR models.

4 Recommendations for Building Robust QSAR Models

In practice, the development of QSAR models can be carried out to reveal the relationship between the chemical structures and their respective endpoint through the use of various types of mathematical and statistical methods for constructing predictive models that can reveal the origin of bioactivity of interest. A typical $m \times n$ data matrix is comprised of m descriptors and n compounds. A closer look at the M descriptors revealed that it is typically comprised of a set of \mathbf{X}_{ij} descriptors and an \mathbf{y}_i endpoint. In a nutshell, a typical QSAR model is essentially described by an equation the form of $\mathbf{Y} = f(\mathbf{X}) + error$ that can be used to predict the endpoint for new compounds in lieu of cost and time-consuming approaches. The classical QSAR modeling workflow can be broken down into five prime steps as demonstrated in Fig. 1.

Thus far, several thousands of QSAR models have been developed for various endpoints and these models are created using different model construction schemes (e.g. stringency of data pre-processing, descriptor types, learning methods and evaluation metrics) and published in the public domain (i.e. this is not including the thousands of QSAR models developed in pharmaceutical companies that are not ever published). The variability in the methods used for the QSAR models and their quality may obviously give rise to different outcome for the conclusions possible to draw from them. To further complicate the picture, the reproduction of QSAR models by following the often rather vague instructions in the Methodology sections of research articles do not always yield the same outcome as in the original article owing to the aforementioned factors.

Fig. 1 General workflow of QSAR modeling. Raw data compiled from the literature or public databases are often noisy and dirty and therefore requires curation to clean the data. In this example, redundant chemical structure is removed followed by descriptor calculation, model building and model performance evaluation

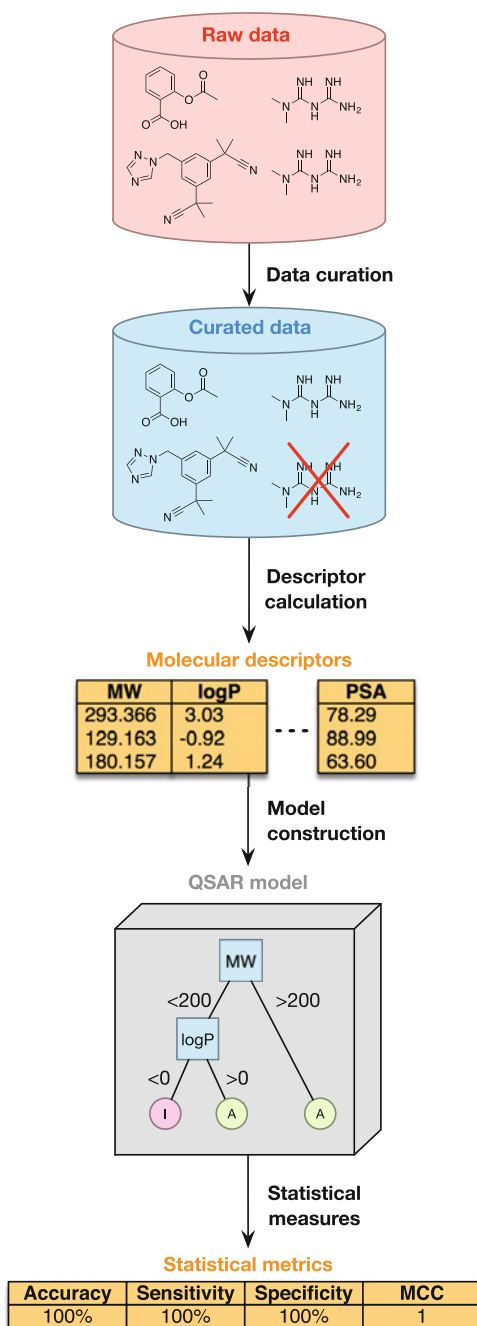


Table 1 Summary of the OECD principles for QSAR modeling

No.	OECD principles	Description
1	Defined endpoint	To ensure that all endpoint values within a given data set are consistent
2	Unambiguous algorithm	To ensure transparency and reproducibility of the proposed QSAR model
3	Defined applicability domain	To determine the boundaries in which the model is robust for predicting query compounds
4	Measures of model's predictive potential	To evaluate the internal and external predictive power of the model
5	Mechanistic interpretation	To ensure that the underlying mechanism of action of compounds can be elucidated

Thus, owing to such lack of standards in QSAR/QSPR modeling, the OECD principles was established to address such issues. This first draft initially took place in Setubal, Portugal in 2002 and a revised version in Paris, France in 2004 at the *Workshop on Regulatory Acceptance of QSAR Modelling for Human Health and Environmental Endpoints* and *37th Joint Meeting of Chemicals Committee and Working Party on Chemicals, Pesticides & Biotechnology*, respectively (Worth and Cronin 2004). It has been mandated that to facilitate the consideration of a QSAR model for regulatory purposes, the model should conform to the five principles summarized in Table 1.

Moreover, the integrity of a QSAR model could be pursued by following suggested sets of standards and best practices (Dearden et al. 2009; Tropsha 2010; Tropsha et al. 2003; Dimova and Bajorath 2016; Spjuth et al. 2010) in the development of robust QSAR models. Particularly, Tropsha et al. stressed the importance of leave-many-out validation, bootstrapping, Y-scrambling test and external validation. Moreover, conflicting viewpoints exist on whether to evaluate the robustness of QSAR models on the basis of external validation in which Hawkins et al. (2003) is against this while Esbensen and Geladi (2010) is in support of this. Moreover, recent investigations clearly favor cross-validation over a single external one (Gütlein et al. 2013; Rácz et al. 2015).

In a nutshell, the development of robust QSAR models should address the following key issues:

1. *Data curation*—The curation or pre-processing of data sets prior to performing any form of data analysis is of utmost importance for QSAR modeling. Raw data sets are often *noisy* or *dirty* in the sense that they may inherently contain redundant compounds, redundant descriptors, incorrect representation of the chemical structure or molecular charge. Curation helps to *clean* and increases the reliability of the data set for subsequent analysis.

2. *Modelability*—Modelability is an a priori estimate of the feasibility to obtain externally predictive QSAR models. Modelability is based on the fact that QSAR models are influenced either by data set characteristics (i.e. size, chemical diversity, activity distribution, presence of activity cliffs, etc.), or by modeling workflow steps (e.g. data set curation, feature selection, external validation, consensus modeling, applicability domain, etc.). Particularly, influences arising from the composition of the modeling workflow can be quantified and can be varied given the wide range of molecular descriptors and machine learning methods that are available. However, effects of data set characteristics can be rather difficult to quantify. While size and chemical diversity are subjective attributes of a data set and are difficult to quantify, recent advances have provided methods for objective quantification of activity cliffs (Guha and Drie 2008; Seebeck et al. 2011; Bajorath 2014; Stumpfe et al. 2014; Hu et al. 2012). Building on the earlier proposed concept of the activity cliffs, Golbraikh et al. (2014) proposed a novel modelability index (MODI) that can be easily computed for any dataset at the onset of any QSAR investigation.
3. *Reproducibility*—This important issue is often overlooked by the QSAR community. This is particularly true as often times, QSAR models are built using proprietary software or code that are often restricted to a selected few and not accessible to the general public thereby precluding further attempts to make use of these models. Moreover, the reproduction of QSAR models is a very difficult task indeed as the construction of QSAR models employs different data sets (e.g. different version of the same bioactivity databases such as ChEMBL 19, 20 or 21; it is also highly likely that data sets focused on the same target protein and performed by different laboratory tend to contain different compounds as they may be compiled from different papers), descriptor types, learning methods and evaluation metrics. Spjuth et al. (2010) examines this issue by proposing an open XML format known as QSAR-ML to formalize QSAR data sets with meta-data, which will facilitate the exchange and reproducibility of the model.
4. *Model validation*—The robustness of QSAR models is reliant on stringent validation of QSAR models. Several validation strategies including (1) randomization of the modelled property also known as Y-scrambling, (2) k -fold cross-validations and (3) external validation using rational division of a data set into training and test sets are currently the de facto standard for ensuring the utility of a model for virtual screening (Tropsha et al. 2003).
5. *Outliers*—Outlying compounds are those molecules which have unexpected biological activity and do not fit in a QSAR model owing to the fact that such compounds may be acting in a different mechanism or interact with its respective target molecules in different modes (Nantasenamat et al. 2009; Verma and Hansch 2005). Similarly, conformational flexibility of target protein binding site (Kim 2007a) and unusual binding mode are attributed as the possible source of outliers (Kim 2007b). Mathematically speaking, an outlier is essentially a data point that has high standardized residual in absolute value when compared to the other samples of the data set. Furthermore, the building of robust and reliable QSAR models generally emphasizes two major aspects: (1) feature selection and

- (2) outlier detection. The two problems are interrelated as outlier definitions are dependent on the selected features. In the realm of QSAR, outliers can be classified as belonging to the following two types: (1) those that fall outside the applicability domain or (2) activity cliffs as discussed in the next section. As the applicability domain considers both chemical and biological space, therefore outliers with respect to biological space can be safely eliminated from QSAR models. However outliers defined based on the chemical space needs further attention. Recent methods such as those from Cao et al. (2011) have argued in support for simultaneously performing variable subset selection and outlier detection using the idea of statistical distribution that can be simulated by the establishment of many cross-predictive linear models. Their approaches build on the concept that the distribution of linear model coefficients provides a mechanism for ranking and interpreting the effects of variables while the distribution of prediction errors provides a mechanism for differentiating the outliers from normal samples (Cao et al. 2011).
6. *Applicability domain*—The applicability domain (AD) (Sahigara et al. 2013) of a QSAR model defines the model limitations with respect to its structural subspace and response space. AD is an indication of the degree of generalization of a given predictive model. AD associated with an endpoint prediction is often well defined if the endpoint prediction for a chemical structure is within the scope of the model. The AD is thus critically reliant on the sampling of chemical subspace and the range of biological readouts that are used for the model development (Sheridan 2015). A commonly overlooked aspect in AD is also the influence of molecular descriptors, generally degenerate and transparent molecular descriptors such as $\log P$, pK_a , etc. afford better degree of generalization to the model while lacking the superior predictive abilities of the more recent topological graph-based descriptors. The various approaches for AD determination are classified as range-based (e.g. bounding box, principal component analysis bounding box and convex hull) and geometric methods (e.g. k -nearest neighbours, DTs, probability density based methods) (Sahigara et al. 2012).
 7. *Structure-activity cliffs*—Compounds within a congeneric series whose subtle differences in the chemical structure lead to striking differences in the observed bioactivity are called activity cliffs (Bajorath 2014). Although, the activity cliffs are appealing to medicinal chemists their presence may be detrimental to QSAR models. The inclusion should be carefully reviewed after analyzing for filters such as PAINS (Baell and Holloway 2010) as unusual activity could be due to a wide range of mechanisms such as outliers of different kinds or even the presence of reactive functional groups (Saxena and Prathipati 2006). However, these compounds belonging to the *activity cliffs* are currently categorized as outliers and frequently removed from QSAR models (Guha and Drie 2008). The MODI quantifies the extent of activity cliffs and serves as a guide to the modelability of a data set (Golbraikh et al. 2014).
 8. *Feature selection*—The number of molecular descriptors that can capture various aspects of a chemical structure have proliferated in recent years (Todeschini and Consonni 2008). Hence, feature or variable selection is an important and hot

area of research (Guyon 2003; Eklund et al. 2014; Goodarzi et al. 2013). In the context of QSAR studies, feature selection improves interpretability by neglecting non-significant effects thereby reducing noise, enhancing generalization by reducing overfitting (also known as reduction of variance), increasing the models' predictive ability and speeds up the QSAR model building process (Saxena and Prathipati 2003). Some widely used and relevant approaches for QSAR studies includes: (1) all subset models (ASM), (2) sequential search (SS), (3) stepwise methods (SW), (4) genetic algorithm (GA), (5) particle swarm optimization (PSO), (6) ant colony optimization (ACO), (7) least absolute shrinkage and selection operator (LASSO), (8) elastic net and (9) variables importance on PLS projections (VIP) (Eklund et al. 2014), (10) correlation-based feature selection (CFS) (Hall 1999), (11) simulated annealing (Siedlecki and Sklansky 1988), (12) sequential feature backward selection (Pudil et al. 1994), (13) sequential feature forward selection (Pudil et al. 1994), (14) minimum-redundancy-maximum-relevance (mRMR) (Peng et al. 2005), (15) ReliefF (Liu and Motoda 2007), (16) Tikhonov regularization (Destrero et al. 2009), (17) recursive feature elimination (RFE) (Guyon et al. 2002), (18) random forest (RF) (Breiman 2001), (19) decision tree (DT) (Quinlan 1993), etc.

9. *Class imbalance*—Class imbalance in supervised machine learning is a major confounding problem for the construction of QSAR models (Li et al. 2009). In a classification setting, the size of the active and inactive sets of compounds may be significantly disproportional and may therefore lead to biased predictive models. Several solutions that include artificially undersampling the overrepresented class or oversampling the underrepresented class or using one class learning or cost-sensitive training have all been suggested as possible remedies to address this issue (Zakharov et al. 2014; Capuzzi et al. 2016).
10. *Chance correlation*—Objectivity is a critical component of any hypothesis generating workflow including QSAR. It has been stressed that causation and correlation are indeed two different things and that a model's performance may possibly arise by chance. A possible remedy is to apply Y-scrambling (Rucker et al. 2007) to evaluate model robustness.
11. *Confidence/reliability of the model*—QSAR models are not universally applicable as predictions may fail under certain conditions. QSAR models are based on mathematical formulations for modeling the bioactivity as well as to draw conclusions from. Their utilization in medicinal chemistry encompasses idea generation, virtual screening and knowledge discovery. Hence, the confidence in the predictions derived from QSAR model should be accessible. Substantial efforts have been devoted to research on this topic within the QSAR community over the last decade and a number of methods have been suggested for estimating the confidence of QSAR predictions. These confidence estimates are typically based on the very loosely defined concept of a QSAR models applicability domain (AD), which is described as the response and chemical structure space in which the model makes predictions with a given reliability. The assumption is that the further away a molecule is from a QSAR models AD, the less reliable the prediction becomes. This confidence measure can be afforded by an approach

known as conformal prediction (Shafer et al. 2008), which has been successfully applied in QSAR modeling (Eklund et al. 2012). The conformal prediction framework provides a unified view of the different approaches for estimating a QSAR models AD. Moreover, conformal prediction provides a natural and intuitive way of interpreting the AD estimates as prediction intervals with a given confidence.

12. *Interpretability of the model*—Perhaps, the most important contribution of QSAR modeling lies in their ability to propose a hypotheses to rationalize the binding/function modulation phenomenon via interpretation of the model's features. In view of its critical role in fulfilling the objectives of QSAR modeling, we focus our chapter on their interpretability. The hypothesis gleaned from QSAR models can benefit biologists and chemists by providing insights into the cause-effect relationships between molecular features and bioactivity measures. These insights can aid medicinal chemists to design future SAR studies objectively and comprehensively. They can also assist molecular and structural biologists in proposing candidates for site-directed mutagenesis and related structure-function experiments. This chapter proposes the use of interpretable molecular descriptors together with interpretable machine learning methods. Recent interest in the field had also shifted towards making the black box learning methods more transparent and amenable to interpretations, which will be covered in the forthcoming sections.

5 Trade-Offs Between Performance and Interpretability

Over the past decades, many QSAR studies had predominantly focused on enhancing and improving the predictive performance instead of the interpretability of the model (Fujita and Winkler 2016). The shift can be seen in QSAR model descriptors moving away from the physicochemical and indicator variables of Hansch-Fujita and Free-Wilson approaches towards highly non-degenerate and continuous molecular descriptors which offer high predictive power. However, improved understanding of the concepts of bioisosterism and the molecular recognition events, identification of problems associated with capturing molecular structures and errors in assay data of widely used SAR databases give credence to the use of moderately degenerate and interpretable 1D or fingerprint based molecular descriptors as expanded elsewhere in this chapter. Learning methods in QSAR modeling have evolved from simple interpretable methods such as linear regression as used by Hansch and Fujita to the complex black box approaches such as neural networks and deep learning. While many experts agree with the obvious improvements (i.e. approximately 10%) to the predictive power from these complex machine learning methods, they argue that the loss of interpretability of the feature contributions are not worth the gain in predictive power. Hence, in Sect. 8.1.4 we expand upon the recent advances in rule extraction techniques that help to provide enhanced interpretation of the complex black box approaches. This section also presents several recent enhancements that

significantly improve the predictive power of the white box learning approaches. Hence, this chapter presents and advance the case for interpretable QSAR models in drug discovery research. We argue that a simple and interpretable QSAR model with modest predictive performance would be more valuable to experimental scientists than a highly predictive but black box model since no or minimal insights can be gained from it.

6 Reverse Engineering of QSAR models

Designing new molecules corresponding to the given biological activity is invaluable to the chemical, material and pharmaceutical industries. The traditional approaches of computer-aided molecular design based on QSAR modeling can be used to solve two main problems: (i) *forward QSAR problem*, which identifies the compounds' structural and physicochemical features related to the experimental readout using machine learning (ii) *inverse QSAR problem* that seeks to reconstruct compounds' structures which correspond to the specific features related with the readout (Faulon et al. 2005; Brown et al. 2006).

The inverse problem is generally addressed as a subgraph construction. Previously, there were five types of approaches to solve the inverse problem: random search, heuristic enumeration, mathematical programming, knowledge-based system, and graphical reconstruction methods. The inverse QSAR analysis is quite challenging for various reasons: combinatorial complexity of the search space, design knowledge acquisition difficulties, nonlinear structure property correlations, and problems in incorporating higher level chemical and biological knowledge (Venkatasubramanian et al. 1995). Thus, it is not surprising that constructing new structural compound given a desired activity is a long-standing problem. In practice, the inverse QSAR method can be divided into the common four steps (Skvortsova et al. 1993; Wong and Burkowski 2009; Churchwell et al. 2004; Visco et al. 2002; Weis et al. 2005). Firstly, a QSAR equation is constructed to derive a forward QSAR model that essentially discerns the relationship between a set of descriptors and their activities. The second step is to generate the set of constraint equations with integer coefficients. The constraints are used for ensuring that the constructed compounds afford the desired activities. There are two types of constraint equations: graphical and consistent equations, which are then solved in the third step. Finally, the compound structures are enumerated and constructed to afford the desired activity while their activities are predicted using the forward QSAR model described in the first step.

Until now, there are relatively few studies providing computational-based models for solving this problem (Visco et al. 2002). Almost all of the proposed computational-based methods that are used are essentially a stochastic model in nature and use either genetic algorithm (GA) or Monte Carlo simulated annealing approach to construct new chemical compounds. In 1995, Venkatasubramanian et al. (1995) and Sheridan and Kearsley (1995) proposed a stochastic model based on Monte Carlo. GA is a general purpose approach based on the Darwinian

principle for natural selection and evolution, which are used for stochastic, evolutionary search, and optimization strategies. The main advantage of GA lies in its ability to allow a dynamically evolving population of molecules to gradually improve by competing for the best performance. However, the problem from these studies represent a combinatorial explosion (Kvasnicka and Pospichal 1996). In order to analyze a huge number of compounds, Kvasnicka and Pospichal (1996) developed a new approach based on a random search that not only afford all solutions but also provide users with a high probability of deriving the correct solution. In 2002, Visco et al. introduced the use of signature descriptors to represent compounds as molecular graphs. In this study, a set of 121 HIV-1 protease inhibitors were analyzed by comparing the proposed QSAR model with other descriptor types consisting of connectivity indices, KierHall shape indices, fragments, electrotopological states and information indices. This work also revealed that signature descriptors are particularly well suited for tackling the inverse problem (also see the work from Faulon 1994, 1996; Faulon et al. 2003; Churchwell et al. 2004; Faulon et al. 2004; Weis et al. 2005). Also from the same group, Churchwell et al. (2004) applied the inverse QSAR approach to a small set of peptide inhibitors that targets the leukocyte functional antigen-1 (LFA-1)/intercellular adhesion molecule-1 (ICAM-1) complex. Their prediction results showed that the predicted IC_{50} values were very close to that of the experimental IC_{50} values. Practically, the inverse QSAR problem is relatively difficult when compared to the forward QSAR problem because the molecular descriptors used for constructing the inverse QSAR model must adequately address the forward QSAR model for the activity or property of a given data, if the subsequent recovery phase is to be meaningful. Additionally, a major problem is to reconstruct and enumerate the chemical structures from its extracted descriptors. To solve such problem, Wong and Burkowski proposed (Wong and Burkowski 2009) a new workflow using a vector space model molecular descriptor (VSMMMD) to represent the chemical structures. Their proposed inverse QSAR model consists of five key steps: (i) calculating the VSMMMD for each compound from the training set; (ii) apply the kernel function (i.e. more detail is discussed in a subsequent section) to map each VSMMMD from the input space (i.e. low dimension) to the feature space (i.e. high dimension); (iii) designing a new point in the feature space using a kernel function algorithm; (iv) map the new point from the feature space and trace back to the input space using a pre-image approximation algorithm and (v) building the chemical structures using the VSMMMD recovery algorithm.

As can be seen, inverse QSAR models has great potential for obtaining desirable compounds directly from the trained QSAR model. Further work in this area is highly encouraged as to help steer towards the practical utility of QSAR models for building promising chemical structures aside from making predictions of their bioactivity values or class label.

7 Interpretable Molecular Descriptors

7.1 Role of Molecular Descriptors in Post-genomic Drug Discovery

Molecular descriptors encode the physical and chemical properties of molecules of interest and are central to QSAR/QSPR studies (Danishuddin 2016). The availability and the use of high quality, interpretable descriptors can greatly contribute to the formulation of an intuitive model for retrospective and prospective analysis of life or material sciences data (Cherkasov et al. 2014). As depicted in Fig. 2, molecular descriptors play a critical role in enabling mathematical and statistical analysis for relating chemical structure with biological data. While human intuitive molecular graphics depictions use the atom, bond, angle coordinates together with charge

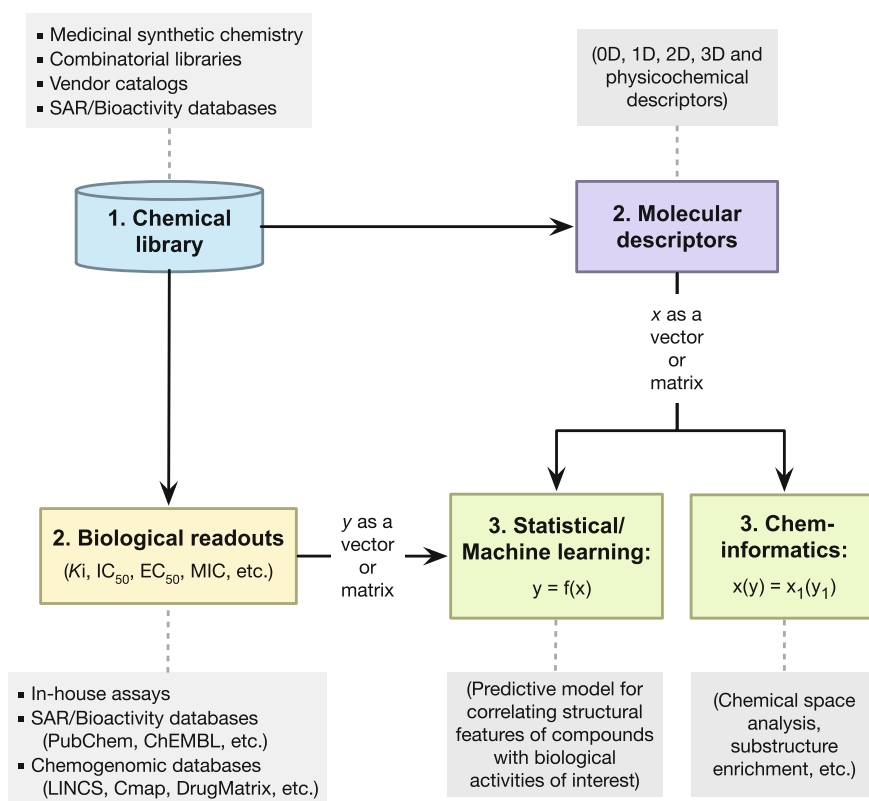


Fig. 2 General schematic diagram depicting the importance of molecular descriptors for capturing the details of chemical structures (from a chemical library) as vectors and matrices; hence enabling mathematical and statistical procedures for QSAR and other chemoinformatics analysis

information to reconstruct the chemical structures as 2D and 3D projections, encoding chemical structure as machine readable matrices and vectors is required for performing mathematical and statistical analysis. In this regard molecular descriptors play a key role in establishing QSARs and in performing chemo-informatics tasks such as chemical space mapping, substructure analysis, etc. In the pre-genomic era, biological readouts were available as single vectors, however advances in miniaturization, robotics and automation in the post-genomic era presented QSAR researchers with a complex array of biological data as matrices. The complex biological matrices include both the traditional target and phenotypic measurements and the recent clinical chemistry and histopathology findings and microarray and proteomics data (Prathipati and Mizuguchi 2016a). These data were generated in standardized high-throughput format and are available in databases such as LINCS, Open TG-Gates, CEBS, DrugMatrix and CMap (Prathipati and Mizuguchi 2016a). Several advanced multi-label statistical techniques (such as network-based inference) and complex molecular descriptors (such as proteo-chemometric) are presently under development which can capture both the biological data's relationship with the chemical structure together with complex relationships among the biological readouts and the chemical structures (Prathipati and Mizuguchi 2016a). Thus a range of machine learning methods are under consideration for multi-label QSAR models depending on the data types such as support vector machines (SVMs), neural networks (NN), k -nearest neighbors (k NN), boosting methods for unrelated multi-label datasets and similarity based approaches such as DT-hybrid, kernel regression methods such as lasso or elastic nets or pairwise kernel method (PKM) for related multi-label datasets (Prathipati and Mizuguchi 2016a). While some of these machine learning methods are discussed in Sect. 8, in the following subsections we expand upon the range of molecular descriptors and their attributes and their utility for modelling the wide array of biological readouts.

7.2 Interpretability of Molecular Descriptors Advances Ligand-Based Approaches

The continuing appeal of QSAR models as part of ligand-based approaches in the face of the ever increasing structural data of target proteins and advancements in structure-based approaches is an interesting conundrum (Prathipati and Mizuguchi 2016a). Although structure-based approaches are highly interpretable and intuitive to drug researchers, their efficiency and effectiveness is limited by several factors including ambiguity in pose prediction, limitations of scoring functions at capturing the molecular recognition event, limitations of existing methods in considering bridging water molecules and induced fit phenomenon (Prathipati et al. 2007; Prathipati and Mizuguchi 2016b). Furthermore, drug targets such as nuclear receptors, G protein-coupled receptors (GPCRs) and kinases are known to have multiple conformational states that exists in equilibrium in the absence of their cognate

ligands (Spyrakis and Cavasotto 2015; Zhao et al. 2014; Rueda et al. 2009, 2010). Most often the X-ray structures of one or the other of these conformational states are difficult to obtain. For instance, several kinases are known to exist in at least 4 different conformational states (e.g. DFG-in, DFG-out, A-loop-out and A-loop-in) in recognizing type -I, -II and -III inhibitors (Chiu et al. 2013). The DGF-out inactive conformational state of a kinase is quite flexible and is quite difficult to crystallize where the catalytically important p-loop is most often difficult to resolve (Kufareva and Abagyan 2008). Similarly, GPCRs too exist in the active, inactive and apo conformational states. While the inactive GPCR conformational states are easy to crystallize owing to its rigidity as conferred by the strong salt-bridge interactions between the helices (e.g. helices 3 and 6 and helices 2 and 5), the active conformational state stabilized in the presence of an agonist disrupts these interactions through charge neutralization, hence becomes flexible and is difficult to crystallize and resolve (Standfuss et al. 2011). Conversely, ligand-based QSAR models are quick and can be dynamically adapted to model both target and phenotypic endpoints as well as different types of chemotypes with relatively little effort (Prathipati and Saxena 2005). QSAR models derived using molecular descriptors were shown to provide high predictive power and were successfully used for hit identification (Krasavin 2015; Geronikaki et al. 2008; Poroikov et al. 2003). The disadvantages of this approach is their comparatively low intuitiveness and their difficulty for interpretation (Saxena and Prathipati 2006). Hence, we shall attempt to discuss the pros and cons of various descriptors in terms of their quality and interpretability.

7.3 Assessing the Quality and Interpretability of a Molecular Descriptor

Historically, the Hammett equation (Hammett 1937) describes one of the earliest known mathematical formulations relating structures with the property of interest (i.e. reactivity in this instance) and remains the most widely used and understood mathematical equation to date. It describes a linear free-energy relationship relating rate or equilibrium of a reaction with a substituent's position and electronic property (i.e. withdrawing or donating) captured as 'Sigma' (Hammett 1937). The molecular descriptor 'Sigma' as proposed by Hammett (1937) to explain the acidity of substituted benzoic acids also serves as useful guidepost in evaluating the quality and interpretability of a molecular descriptor. 'Sigma', also called the substituent constant, has several features that makes it an excellent molecular descriptor, particularly it has (1) high structural interpretation, (2) good correlation with biological or physical property (i.e. pK_a in this case), (3) can be applied to local structure (substructures), (4) uses the familiar structural and electronic concepts (e.g. electronegativity and polarizability), (5) high sensitivity (i.e. varies with structures; even isomers) and (6) size dependence (i.e. changes with molecular weight). However, the original implementation of Hammett involves using experimental properties and makes the

Table 2 Summary of the strengths and weaknesses of the various dimensions of molecular descriptors. The number of stars denotes the strengths and weaknesses for each characteristics while the exclamation mark designate that caution should be taken

Characteristics	0D	1D	2D	3D	PC
Simplicity	***	***	**	*	***
Calculation efficiency	***	***	**	!	*
Structural interpretation	*	**	*	***	***
Correlation with biological property	*	**	**	***	****
Applicable to local structure (substructures)	*	***	**	*	**
Use familiar structural and electronic concepts	*	*	**	***	****
Sensitivity (discriminate different structures including isomers)	!	*	**	***	**
Size dependency (varies with MW)	*	**	**	**	**

0D: zero-dimensional descriptors, 1D: one-dimensional descriptors, 2D: two-dimensional descriptors, 3D: three-dimensional descriptors, PC: physicochemical descriptors

computation of sigma highly inefficient and hence not practical for high-throughput virtual screening workflows. We shall discuss the importance of physicochemical properties descriptors and the 4 major class of structural descriptors in light of the features discussed above (Table 2). Furthermore, several novel applications of QSAR such as the modelling of peptides, nucleotides and nanostructures for biologics-based drug discovery research requires the availability of novel descriptors. Hence, Table 4 presents the list of free software along with availability of various descriptor types (Table 3).

7.4 Trade-Offs Between Descriptor Quality and Interpretability

Thus, it should be noted that a descriptor's quality and its interpretability, together with the use of an appropriate machine learning method can greatly produce a practical and interpretable QSAR model that scientists can use. The *sensitivity* or the *degeneracy* of a molecular descriptor is the measure of its ability to avoid equal values for different molecules. This is the most critical attribute of a descriptor's quality. Furthermore, a descriptor's interpretability can be defined as its ability to elucidate and rationalize the underlying structural and physicochemical properties responsible for the biological response.

3D descriptors which most accurately encode the structural and physicochemical properties that are responsible for the investigated endpoint are presently regarded to afford robust quantitative descriptions of molecular structures. They have high

Table 3 Summary of model techniques used in QSAR modeling and their advantages and disadvantages

Method ^a	Interpretable	Linear	Supervised learning?	Advantage(s)	Disadvantage(s)
MLR	Yes	Yes	Yes	Good interpretability	Problem of learning dichotomous variables
LR	Yes	Yes	Yes	Deal with dichotomous variables	Perform poor on complex data
ELM	Yes	Yes	Yes	Interpretability	Perform poor on multiclass data
PCA	Yes	Yes	No	Dimension reduction	Unsupervised learning
PLSR	Yes	Yes	Yes	Interpretable and reduced dimension	Linear model
DT	Yes	No	Yes	High interpretability	Overfitting
RF	Yes	No	Yes	High interpretability/tolerant to overfitting	Long training time
ANN	No	No	Yes	Perform well on complex data	Poor interpretability
DL	No	No	Yes	Hierarchical features learning	Poor interpretability
SVM	No	No	Yes	Good generalization performance	Poor interpretability

^aMLR: multiple linear regression, LR: logistic regression, ELM: efficient learning method, PCA: principal component analysis, PLSR: partial least squares regression, DT: decision tree, RF: random forest, ANN: artificial neural network, DL: deep learning, SVM: support vector machine

sensitivity and present different values of different isomers and other subtle structural variations. Some 3D descriptors such as those based on the GRID concept or obtained from quantum chemical computations provide causal insights while those based on the graph concept akin to the 2D graph-based descriptors present very little causal interpretation. Furthermore, 2D graph-based descriptors are equally as degenerate as a 3D descriptor and can also be regarded as a descriptor of high quality. However, most medicinal chemistry SAR data are not highly sensitive to small changes in the structure (i.e. the addition of substructures to non-pharmacophoric areas) and are shown to have moderate complexity (Schuffenhauer et al. 2006). Furthermore, the assay data too are prone to experimental artifacts (e.g. aggregation, reactive functional groups induced assay readouts) and errors (i.e. standard deviation of technical replicates) (Feng et al. 2005; Feng and Shoichet 2006; Feng et al. 2007; McGovern et al. 2002; Thorne et al. 2010). The moderate complexity of the chemical space can be attributed to the difficulties in their synthesis and purification as well as the characterization of stereo- and regioisomers.

In light of the moderately complex chemical space, 1D or fingerprint descriptors having moderate sensitivity (e.g. non-degenerativity) and interpretability, have become the de facto standard in chemoinformatics both for a prospective and retrospective QSAR analysis (Schuffenhauer et al. 2006). The compact nature of the bit-vector representation makes them amenable to not only QSAR modeling but also for a wide range of computations such as similarity searching (Prathipati et al. 2008), clustering (Prathipati et al. 2008), substructure searching and the inverse QSAR problems (Rosenbaum et al. 2011). As to address issues such as assay errors, artifacts and heterogeneity of assay methods, the use of classification models has been proposed as a promising solution and as such its usage has steadily increased in recent years.

7.5 *Dimensions of Molecular Descriptors*

7.5.1 0D Descriptors

The 0D descriptors (Todeschini and Consonni 2008) capture the counts of atoms (e.g. number of carbon atoms, number of nitrogen atoms, etc.) and bonds as well as their constitution (e.g. hybridization states and bond orders). In addition, 0D descriptors also encode the sum or average of the atomic properties such as weight, volume, polarizability, electronegativity, etc. These descriptors are easily calculated and naturally interpreted but they may not be very sensitive to subtle changes in molecular structures (e.g. isoforms). However, this class of descriptors have successfully been used in explaining the variation effect of structures on activity/property of several data sets as has extensively been shown by the research group of Andrey Toropov and Alla Toropova (Toropov and Benfenati 2007a, b; Toropov et al. 2010).

Particularly, the research group of Toropov and Toropova proposed the SMILES-based descriptors for the easy computation and interpretation of the importance of

features followed by QSAR modeling using the Monte Carlo approach. This computational methodology has been produced as a free software called the CORrelation And Logic (CORAL) (<http://www.insilico.eu/coral>) (Toropov and Benfenati 2007a, b; Toropov et al. 2010). The SMILES notation is used to directly extract 1D molecular features (e.g. atom, bond and other elements) from the chemical structures without the need for external software for descriptor calculation. It can be used for the development of regression and classification based predictive models using the Monte Carlo technique for biological activities (Worachartcheewan et al. 2015; Masand et al. 2014), chemical properties (Toropova and Toropov 2014; Gobbi et al. 2016) and nanomaterial properties (Toropov et al. 2013). CORAL requires an input file consisting of the compound name, SMILES notation and the bioactivity values or class labels. Compounds from the data set are separated into training, invisible training, calibration sets (i.e. used as visible data set) and validation set (i.e. used as invisible data set that is not used during the model construction). Moreover, such data subsets are generated for three or more independent data splits as to evaluate variability from the prediction models. The performance of such models can be derived from statistical parameters such as R^2 , Q^2 , $R^2 - Q^2$ (Worachartcheewan et al. 2014).

A set of local and global molecular features can be derived from the SMILES notations as follows:

$$\begin{aligned} abcdef &\rightarrow a + b + c + d + e + f(S_k) \\ abcdef &\rightarrow ab + bc + cd + de + ef(SS_k) \\ abcdef &\rightarrow abc + bcd + cde + def(SSS_k) \end{aligned} \quad (1)$$

These are the examples of local descriptors that represents the elements in the SMILES notation. In addition, global descriptors are also encoded designated as *BOND*, *PAIR*, *NOSP* and *HALO* as follows:

- *BOND* is presence/absence of bond in the SMILES input such as double bond (=), triple bond (#) and stereo chemical bond (@)
- *PAIR* is the co-occurrence of two elements of the following: F, Cl, Br, I, N, O, S, P, #, = and @
- *NOSP* is presence/absence of N, O, S and P
- *HALO* is presence/absence of halogens

In the software, optimized parameters include threshold and correlation weights (CW). An example of equation of SMILES-based optimal attributes, was calculated by the following equation:

$$\begin{aligned} DCW(Threshold, N_{epoch}) &= \sum CW(S_k) + \sum CW(SS_k) + \sum CW(SSS_k) + \\ &\sum CW(BOND) + \sum CW(NOSP) + \sum CW(HALO) + \sum CW(PAIR) \end{aligned} \quad (2)$$

The biological/chemical endpoint can be calculated as follows:

$$Endpoint = C_0 + C_1 \times DCW(Threshold, N_{epoch}) \quad (3)$$

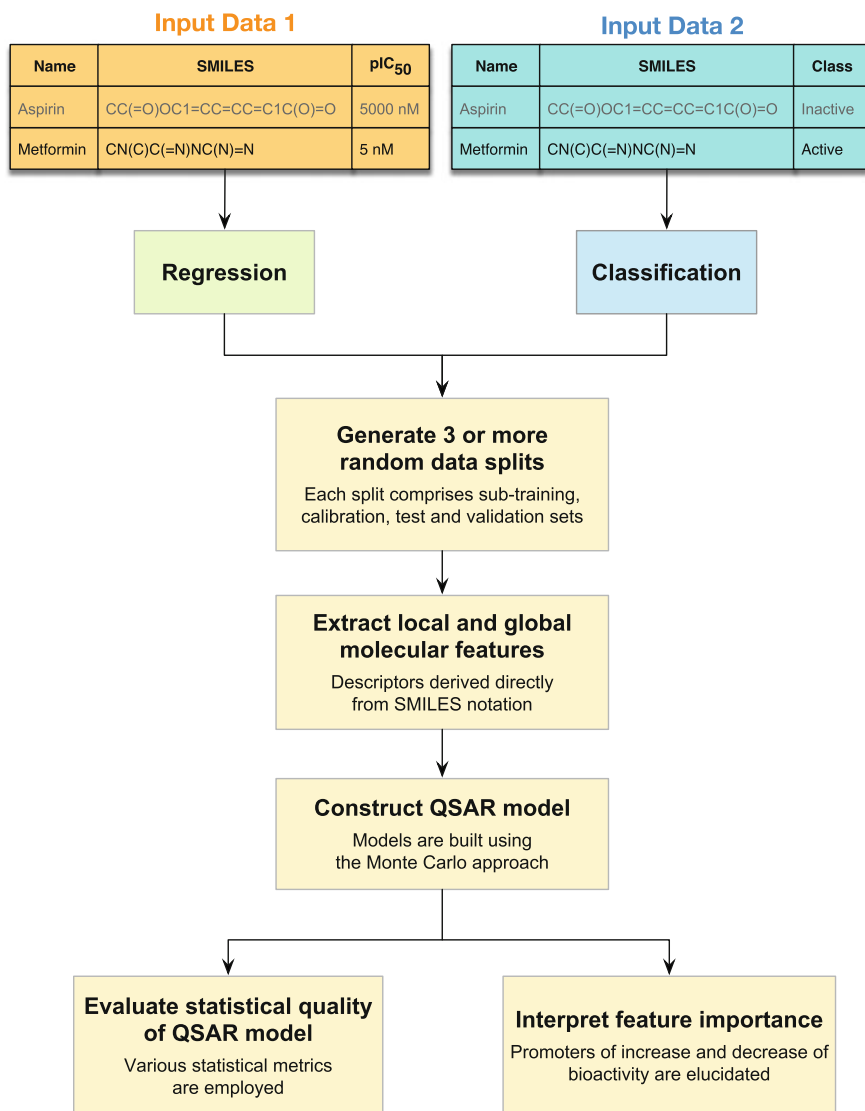


Fig. 3 Workflow of the CORAL software for constructing QSAR modelings using SMILES-based descriptors

where C_0 is the intercept and C_1 is the slope or correlation coefficient.

Furthermore, the molecular fragments obtained from the software can give knowledge of important chemical feature influencing their activities as promoters for increasing or decreasing biological activity. The summary of development of predictive models using SMILES-based descriptors by CORAL software are outlined in Fig. 3.

Recently, Filimonov et al. (2009) proposed a novel QNA-based Star Track QSAR approach in which any molecule is represented as a set of points in 2D space of QNA descriptors. The Star Track approach is in contrast with the classical QSAR method and does not require the use of feature selection. This approach is implemented in the GUSAR software package and is based on a self-consistent regression, QNA descriptors and the topological length and volume of a molecule. This approach predicts quantitative values of biological activity of compounds on the basis of their structural formula and does not require the use of information about the 3D structures of ligands and/or target proteins. The Star Track QSAR approach compares favorably with different 3D and 2D QSAR methods on various gold standard data sets and does not select models based on Q^2 values. Thus, the Star Track QSAR approach as implemented in the GUSAR software package is a potentially useful approach for the derivation of statistically robust, interpretable and fast QSAR models.

7.5.2 1D Descriptors

1D descriptors, also referred to as fingerprints, essentially capture the counts and properties of functional groups and substructural fragments (Todeschini and Consonni 2008). A fundamental difference between 1D descriptors and fingerprints is that the former uses a predefined set of keys (i.e. functional groups and substructures) to generate the descriptors while the latter uses either a predefined set or a set of keys generated on the fly. The older generation of fingerprints consisting of MACCS (Durant et al. 2002), PubChem, and SMARTS still uses a predefined set of keys (Hinselmann et al. 2011) for generating fingerprints and are critically limited at capturing the domain (target- and ligand-) specific structural features responsible for variation in activities. For instance, predefined fingerprints may capture too few or too many correlating features which may have moderate value in QSAR studies. However, recent advances in computer science led to the concept of hashed fingerprints where a set of patterns are generated by gathering atom environment information or subgraph information or both. The generated context dependent patterns are then transformed into hash codes (i.e. a fixed size vector) using hashing algorithm. These hash codes can then be transformed into bit strings using a random number generation of a defined length (i.e. size of the fingerprint). The presence and absence of a pattern is marked as being either 1 and 0, respectively. Extended connectivity fingerprint (ECFP) (Rogers and Hahn 2010) is a prototypical example of a hashed fingerprint. A major advantage of 1D descriptors or hashed fingerprints is their ability to capture complex structural patterns in uniform fixed bit vectors, which can be quickly computed (Rogers and Hahn 2010). These bit vectors are amenable for molecular similarity/substructure analysis problems, show little degeneracy, are naturally interpreted and are widely used in chemoinformatics (Prathipati et al. 2008). In view of the intuitive concepts of substructures' and functional groups' contributions to drug design and their efficient computation, the 1D descriptors or fingerprints were primarily used for the inverse QSAR problems (Rosenbaum et al. 2011) as discussed in the Introduction.

7.5.3 2D Descriptors

2D or topological descriptors (Gozalbes et al. 2002) are computed by encoding the atoms and their connectivity as a graph. Several variations to the graph-theoretic representation of atoms and their connectivities led to the wide plethora of methods for the generation of ‘graph-theoretic’ descriptors such as Kier and Hall (1976), Broto et al. (1984), Balaban (1982), Randic (1975), MEDV etc. Although they lack in interpretability, 2D descriptors can be considered good descriptors in many aspects (as listed in Table 2). However, the poor interpretability of this class of descriptor critically limits its usage in retrospective QSAR analysis (Gozalbes et al. 2002). Furthermore, since correlation does not always imply causality, models derived using these class of descriptors are difficult to prioritize from a pool of models that offer very similar statistical significance (Saxena and Prathipati 2006). There are two excellent techniques to mitigate this problem and discriminate seemingly equivalent models via the generalized pairwise correlation method (GPCM) (Héberger and Rajkó 2002) and the sum of ranking differences (Heberger and Skrbic 2012). However, QSAR models derived from these descriptors are ideally suited for a prospective virtual screening analysis as they can be efficiently computed and generally have very low levels of degeneracy (Saxena and Prathipati 2006).

Among the various topological indices, the molecular electronegativity distance vector based on 13 atomic types called the MEDV-13, is a fast, easy to use, reproducible and predictable descriptor for QSAR studies. The studies by Liu et al. (2001) show the performance of MEDV-13 models were comparable to 3D QSAR studies and are also applicable to QSARs of peptides. MEDV-13 descriptor in addition employs information about an element atom type, valence electronic state, and chemical bond type from 2D molecular topology and requires no information related to 3D structures or physicochemical properties or molecular alignments.

7.5.4 3D Descriptors

3D descriptors characterize the 3D structure of a molecule in terms of their shape, steric and electronic features (Kubinyi 1993). While shape-based 3D descriptors (e.g. volume, *RDF* (Gonzlez et al. 2005), *autocorrelation3D* (Sliwoski et al. 2016), etc.) are highly relevant in explaining SAR data, they remain difficult to interpret. Furthermore, the 3D descriptors comprising of *RDF* (Gonzlez et al. 2005), *3D-MoRSE* (Devinyak et al. 2014), *WHIM* (Bravi et al. 1997) and *GETAWAY* (Consonni et al. 2002) descriptors share many similarities with 2D descriptors as described above. While the latter encodes atoms and their connectivity as simple graphs, the 3D shape-based descriptors capture these features together with their distances and angles as part of a complex graphs. On the other end, the 3D descriptor spectrum includes descriptors such as steric and electrostatic fields that are computed using semi-empirical quantum chemical methods as part of the GRID concept (Sippl 2006). The 3D QSAR paradigm asserts the importance of conformational preferences of compounds for molecular recognition to its target protein in addition to structural

and physicochemical features as described above. The CoMFA/CoMSIA methods (Cramer et al. 1988) to date remains the prototypical examples of this paradigm and several leading publications reported seemingly interpretable retrospective analysis of both target-based (Prathipati et al. 2005) and phenotype-based SAR data. However, in a seminal paper, Doweyko (2004) debunked the commonly asserted illusion and showed that the so-called significant regions are subject to the vagaries of alignment and that the nature of possible interactions heavily depends on the eye of the beholder. Furthermore, the arbitrary nature of both the alignment paradigm and atom description lends itself to capricious models, which in turn can lead to distorted conclusions (Doweyko 2004). In spite of limitations of the 3D QSAR approach, this class of descriptors demonstrates very low levels of degeneracy (i.e. extremely sensitive to changes in the structure) and is considered as the gold standard amongst the QSAR modelling techniques. Although, the 3D steric and electrostatic fields have been very intuitive both for explaining the SAR data and for guiding several novel designs, a potential limitation is their rationalization is limited to a congeneric series of compounds. Hence, 3D-QSAR models are not typically used for large-scale prospective virtual screening analysis (Doweyko 2004). Although, several variations of the Tripos CoMFA/CoMSIA (Cramer et al. 1988) have emerged in recent years, the only known freeware is Open3DQSAR (Tosco et al. 2011), which is potentially an interesting addition to the growing number of 3D QSAR software.

7.5.5 Physicochemical Properties

Physicochemical properties are considered to be one of the most relevant descriptors for drug design (Brustle et al. 2002; Taskinen and Yliruusi 2003). While they are mostly measured quantities, they are calculated based on parameterization with measured data. Thus, these descriptors differ from others in that they are not derived from first principles but are obtained from models trained using either 0D, 1D, 2D, 3D (e.g. 3D quantum chemical descriptors calculated using the GRID approach) to fit with experimentally obtained physical and chemical properties such as $\log P$, pK_a and solubility measures (Taskinen and Yliruusi 2003). Hence, in contrast to some molecular descriptor software and reviews, which had categorized this class of molecular descriptors as 0D, 1D, 2D or 3D. Thus, in this chapter we have placed this class of descriptors separately. These descriptors (e.g. $\log P$, $\log D$, pK_a) play a major role in both pharmacodynamic and pharmacokinetic properties of compounds (Taskinen and Yliruusi 2003). Furthermore, they have now become a part of the standard checklist for assessing the drug-likeness (e.g. Lipinski's rule-of-five) and other pharmacokinetic liabilities. Moreover, they are also widely used in explaining the variation of target-based SAR data. Most proteins' structure-function modulation is mediated via salt-bridges and small molecules typically modulate the function of a protein via charge neutralization thereby leading to the disruption of salt-bridges followed by a consequent change in the structure and function of the protein (Prathipati and Saxena 2005). In this context, physicochemical properties like pK_a and other quantum chemically derived electronic properties are widely used (Manallack 2008). In

spite of their widespread usage, intuitive appeal and interpretability, these descriptors remain difficult to compute. Given the importance of modelling various electronic effects (e.g. inductive, mesomeric, polar) (Thorner 1979; Patani and LaVoie 1996; Jelfs et al. 2007; O’Boyle et al. 2017b; Harding et al. 2009; Morgenthaler et al. 2007; Xing et al. 2003; Manallack 2008), it should be noted that computationally-expensive quantum chemical descriptors are often used to train models that can predict the pK_a , polarizability, etc. Thus, the development of software for computing these descriptors is an area of active research. Improvements in GPU technology have greatly accelerated the utilization of quantum chemical simulations (Patani and LaVoie 1996) for the prediction of physicochemical properties and biological activities.

8 Interpretable Learning Algorithms

8.1 Black Box Learning Methods

Kurgan et al. (2009) used the term black box models to describe the fact that machine learning models do not identify the underlying associations of individual features with the specific outcome as well as not revealing which features provide essential contribution to the observed prediction accuracy. Black box have demonstrated success in modeling a wide range of bioactivities and properties (Charoenkwan et al. 2013; Shoombuatong et al. 2015; Simeon et al. 2016a, b; Shoombuatong et al. 2015; Nantasenamat et al. 2005, 2007a).

8.1.1 Support Vector Machine

Support vector machine (SVM) (Cortes and Vapnik 1995; Burges 1998; Barakat and Bradley 2010) is a statistical learning approach and a well-known maximum margin classifier that is based on the principles of structural risk minimization (SRM). The SRM principle is utilized to seek a hypothesis function with low capacity from a nested sequence of functions that can simultaneously minimize both the true error rate (i.e. prediction error on the external set) and the empirical error rate (i.e. prediction error on the training set) as illustrated in Fig. 4.

Given a training set $D_{Tr}^n = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_m, \mathbf{y}_m)\}$, where $\mathbf{x}_i \in \mathbb{R}^n$ and $\mathbf{y}_i \in -1, +1$, the SVM classifier finds the optimal separating hyperplane that has the largest margin and satisfies the following conditions:

$$\begin{aligned} \mathbf{w}^T \varphi(\mathbf{x}_i) + \mathbf{b} &\geq +1, & \text{for } \mathbf{y}_i = +1 \\ \mathbf{w}^T \varphi(\mathbf{x}_i) + \mathbf{b} &\leq -1, & \text{for } \mathbf{y}_i = -1 \end{aligned} \quad (4)$$

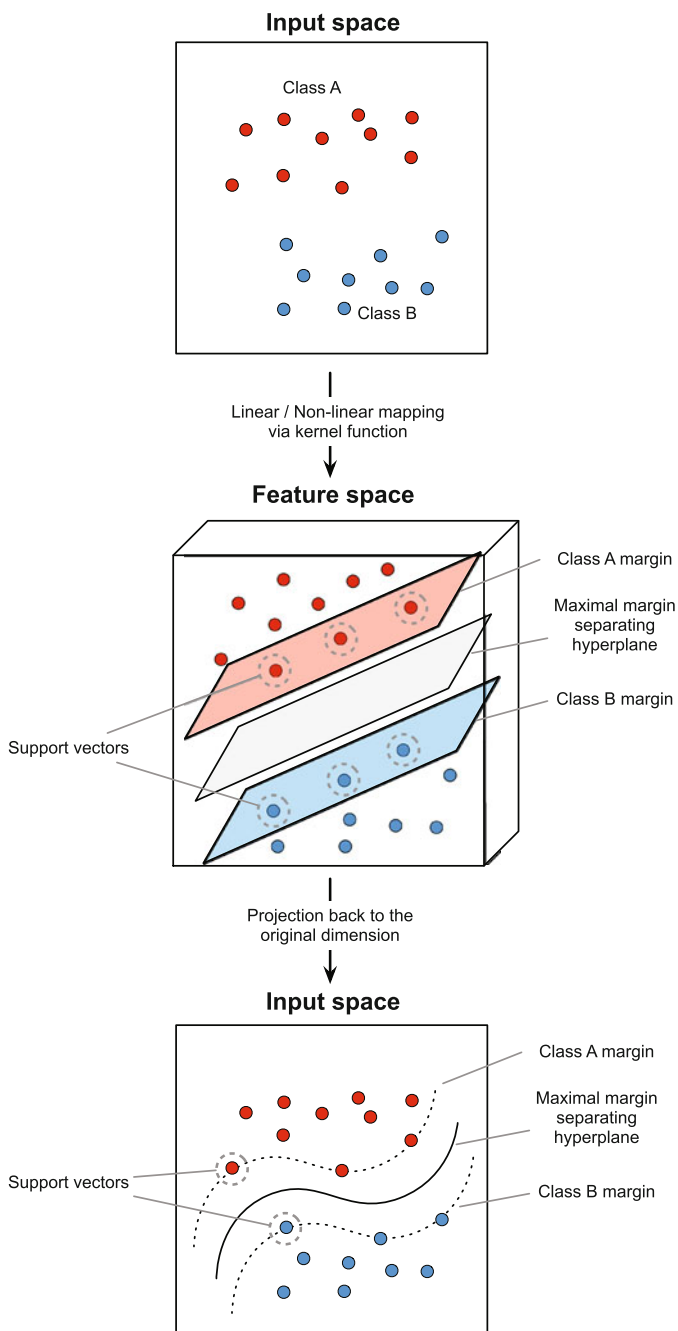


Fig. 4 Illustration of the SVM learning process. Initially, the input space is transformed to a higher dimensional feature space via the use of kernel functions whereby the maximal margin separating hyperplane is obtained after defining the margins of the two classes. It should be noted that compounds (denoted by *circles*) lying on the margin represents the support vectors

which is equivalent to:

$$\mathbf{y}_i[\mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}_i) + \mathbf{b}] \geq +1, \quad i = 1, 2, \dots, m \quad (5)$$

The non-linear function maps the input space to a higher dimensional space called the feature space. The mapping function $\boldsymbol{\varphi}(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}^p$, where $n \ll p$, is performed by defining the inner product between two samples through kernel function $K(\mathbf{x}, \mathbf{y})$. Practically, the kernel function $K(\mathbf{x}, \mathbf{y})$ is expressed with a similarity measurement between two samples in the data set, which is defined as Burges (1998):

$$\begin{aligned} K(\mathbf{x}, \mathbf{y}) &= \boldsymbol{\varphi}(\mathbf{x}) \cdot \boldsymbol{\varphi}(\mathbf{y}) \\ &= \sum_i \varphi(\mathbf{x})_i \varphi(\mathbf{y})_i \end{aligned} \quad (6)$$

For the kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$, the most popular kernel function includes: the linear kernel $\boldsymbol{\varphi}(\mathbf{x}_i)^T \boldsymbol{\varphi}(\mathbf{x}_j)$; the polynomial kernel $(1 + \boldsymbol{\varphi}(\mathbf{x}_i)^T \boldsymbol{\varphi}(\mathbf{x}_j))^d$, where $d = 2, 3$, and 4 (i.e. it should be noted that $d = 1$ for linear kernel); and the radial basis function (RBF) kernel $\exp(-\gamma(\|\mathbf{x}_i - \mathbf{x}_j\|))$, where C (the penalty factor), γ (trading off error predictions against margin width) and ε (the percentage of support vectors in the SVM model) are parameters to be optimized. Kernel functions are often used in SVM because of the scalar product in the dual form. In fact, these approaches can also be used for other machine learning algorithms, but they are not tied to the SVM formalism. It should be noted that the RBF kernel has been widely used in SVM modelling. The decision function of the SVM classifier is given by:

$$y(x) = \text{sign} \left[\sum_{i=1}^m \alpha_i \mathbf{y}_i K(\mathbf{x}_i, \mathbf{x}) + \mathbf{b} \right] \quad (7)$$

where α is the parameter solved by the Lagrangian algorithm and $\mathbf{x} = (x_1, x_2, \dots, x_M)$.

This method was not originally developed as a tool for statistical prediction by Cortes and Vapnik (1995). However, Vapnik enabled the original SVM to solve regression problems also known as support vector regression (SVR), by choosing a suitable cost function (ε -insensitive loss function) that enables a sparse set of support vectors to be obtained. The standard regression procedure is to identify a function $f(x)$ that provides the least square error between predicted and actual observed responses for all training data set. In contrast, SVR attempts to minimize the generalization error bound for achieving higher generalization performance. This generalization error bound is derived from the combination of the training error and a regularization term controlling the complexity of hypothesis space. The first term is calculated by the ε -insensitive losses. The ε -insensitive loss function for SVR method (Drucker et al. 1996; Song et al. 2002) is defined as follows:

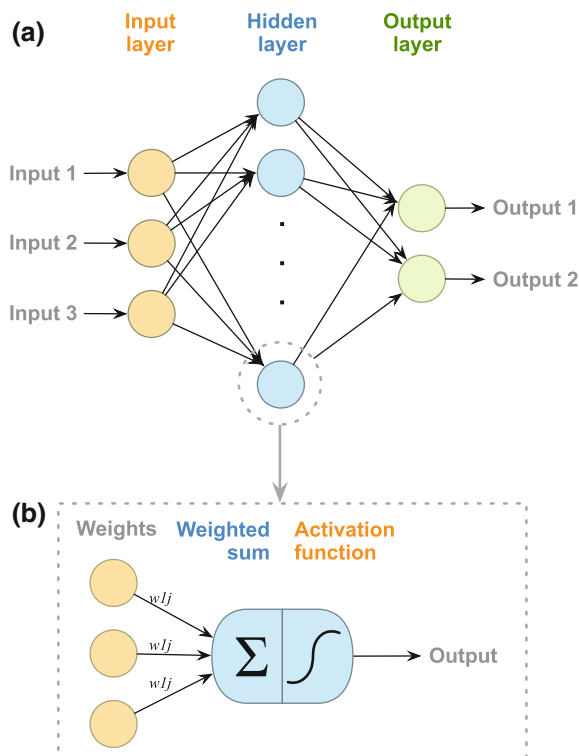
$$L_\varepsilon(\mathbf{y}, f(\mathbf{x}, \boldsymbol{\beta})) = \begin{cases} |\mathbf{y} - f(\mathbf{x}, \boldsymbol{\beta})| - \varepsilon, & |\mathbf{y} - f(\mathbf{x}, \boldsymbol{\beta})| \geq \varepsilon \\ 0, & |\mathbf{y} - f(\mathbf{x}, \boldsymbol{\beta})| < \varepsilon \end{cases} \quad (8)$$

where y is the actual value, $f(\mathbf{x}, \beta)$ is the predicted value (i.e. in which the simple form is $f(\mathbf{x})$, where $\mathbf{x} = (x_1, x_2, \dots, x_n)$) and ε is the insensitivity parameter.

8.1.2 Artificial Neural Network

Artificial neural network (ANN) is a well-established machine learning algorithm for establishing QSAR models (Nantasenamat et al. 2005, 2007a, b, 2008; Worachartcheewan et al. 2009). ANN represents biologically inspired prediction and classification methods whose original development was based on the structure and function of the network of neurons (Zurada 1992). A typical ANN is established with three major components, namely the transfer function, the learning rule and the connection formula (Simpson 1990) as illustrated in Fig. 5. Until now, the feed-forward ANN (FF-ANN) is the most popular ANN that has been used in real-life situation (Ebrahimi et al. 2016). Among many learning algorithm for estimating the parameter of FF-ANN, the back-propagation (BP) algorithm is the most extensively used for finding the optimal parameters, which is carried out by minimizing the error of the network through the derivatives of the error function. For a given training set D_{Tr}^m in a BP-ANN task, the input layer starts to propagate the signal through the connection

Fig. 5 Illustration of the architecture of artificial neural network (a) and inner working of neurons in a hidden layer (b)



weights and the transfer function to produce the output for each neuron. The output or predicted value is then compared to the actual value and the differences in the value between the predicted and actual values is minimized by the BP algorithm. Practically, the delta rule is used to optimize the weights via the BP algorithm:

$$W_{ij}^{new} = W_{ij}^{old} + \Delta W_{ij} \quad (9)$$

$$\Delta W_{ij} = -\mu \frac{\partial E_p}{\partial W_{ij}} out_j \quad (10)$$

where out_j is the output of the j th neuron, μ is the training rate and E_p is the error. The output layer of ANN can be represented mathematically as:

$$O = f\left(\sum_{i=1}^M \mathbf{w}_i \mathbf{x}_i + \mathbf{b}\right) \quad (11)$$

8.1.3 Deep Learning

Owing to the limitations of FF-ANN, a deep learning (DL) method was proposed by three separate groups (Hinton et al. 2006; Raiko 2012; Bengio 2009) for solving the process of training models in many layers. In 2006, DL also known as deep neural network has become increasingly popular for parameter approximation by allowing computational models to learn from representations of data using multiple levels of abstraction (Hinton et al. 2006, 2012). Many research groups reported that there are many different points between ANN and DL (Xing et al. 2003; Leung et al. 2014; Ma et al. 2015). Firstly, each layer of the neural network is constructed from a row of neurons while DL is built from several layers of neurons. Layers in a DL consist of three main layers: (i) the input layer (i.e. the bottom layer), where the descriptors of a molecule are entered; (ii) the output layer (i.e. the top layer), where prediction results are created; (iii) the hidden (middle) layers, where the word “deep” in DL implies that there is more than one hidden layer, as illustrated in Fig. 6. There are two popular choices of activation functions (f) that are used in the hidden (f_H) and output (f_O) layers, namely the sigmoid function and the rectified linear unit (ReLU) function. Secondly, the output layer of ANN basically has one or more neurons and each output neuron generates prediction for a separate endpoint while DL can naturally model multiple endpoints at the same time. Finally, DL employs ReLU instead of sigmoids (i.e. usually used in ANN) as activation functions in order to overcome the vanishing gradient problem. These activation functions have non-vanishing derivative.

Previously, many reports suggested that the predictive performance of DL has dramatically improved as compared to that of standard ANN. The strength of DL lies in its ability to manipulate the intricate structure in large training set by using the backpropagation algorithm. Presently, DL is being applied to many domains of science, business and government. For instance, in the domain of bioinformatics,

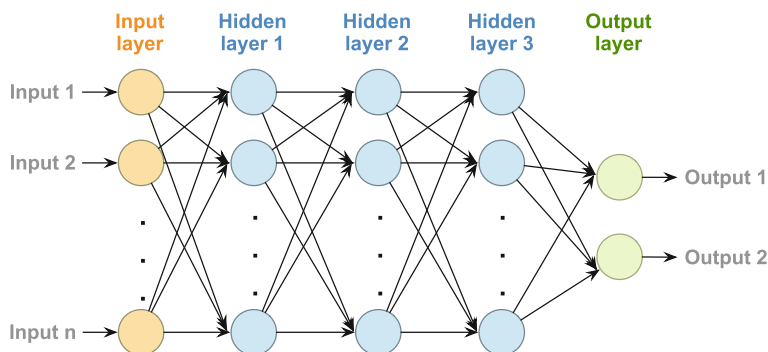


Fig. 6 Illustration of the architecture of deep learning algorithm

DL has been compared with other conventional machine learning algorithm for predicting the activity of potential drug molecules (Ma et al. 2015), analysing particle accelerator data (Ciodaro et al. 2012), reconstructing brain circuits (Helmstaedter et al. 2013) and predicting the effects of mutations in non-coding DNA on gene expression and disease (Xing et al. 2003; Leung et al. 2014). DL has also yielded promising results in natural language processing (NLP) (Collobert et al. 2011), especially for topic classification, sentiment analysis, question answering and language translation (Bordes 2014; Sutskever et al. 2014).

8.1.4 Towards Opening the Black Box

The classical QSAR approach developed by Hansch in the 1960s (Hansch et al. 1962) has a long history in predicting biological activities and physical properties. The original model used a simple, transparent and interpretable MLR model and provided excellent mechanistic interpretation of the biological activity. However, QSAR models are expected to provide both quick predictions (i.e. in a prospective manner) and mechanistic interpretation (i.e. through its features in a retrospective manner). The superior performance of SVM and ANN models vis-a-vis other computational-based models in a variety of application areas is widely known. The high accuracy and robustness of these methods can be attributed to their ability to build non-linear, black-box models that can account for the complexity of the input data. This inability to provide an explanation or comprehensible justification for the predicted solutions critically limits their application to several areas. In application areas such as medical diagnosis, it is highly desirable to give a clear mechanistic interpretation associated with the classification decisions in order to aid the compliance by both the physician and the patient. To mitigate this problem, methods that can aid the interpretation of significant features used by the model can be obtained via the use of rule extraction methods as had recently been shown for ANNs (Fung et al. 2005; Andrews et al. 1995; Setiono et al. 2002) and SVMs (Andrews et al. 1995; Barakat and Bradley

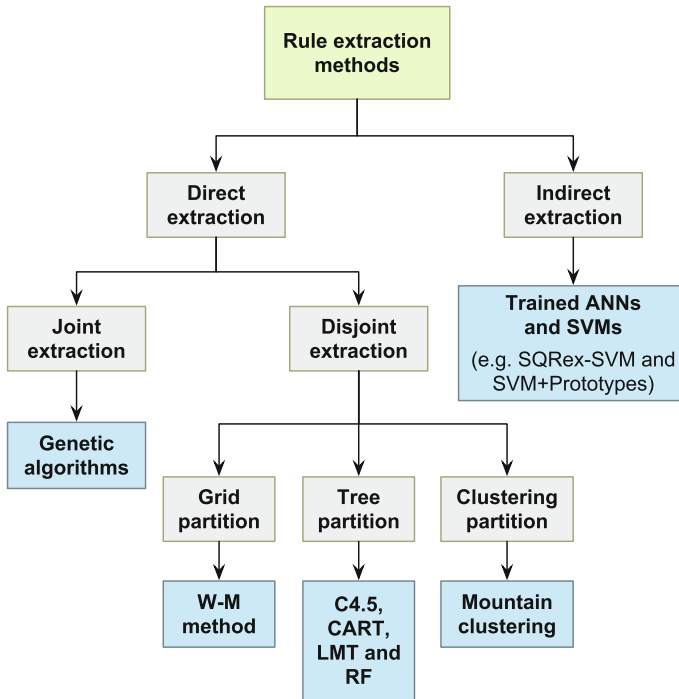


Fig. 7 Taxonomy of rule extraction techniques

2010; Núñez et al. 2002; Zhang et al. 2005; Fu et al. 2004; Barakat and Diederich 2004, 2005).

In recent years, many rule extraction techniques were developed to extract easy-to-understand regularities from data. Figure 7 illustrates the taxonomy of those methods that are derived from the data mining research community. Firstly, they are divided into direct and indirect methods according to the approach that rules are reasoned out. As mentioned, indirect rule extraction methods (e.g. SQReX-SVM and SVM+Prototypes) have been developed for providing explanations as well as affording prediction. Direct rule extraction methods are more widely studied in theory and applied in practice. The direct extraction of rules contains two critical tasks namely antecedent (i.e. representing the condition part of rules) and consequent (i.e. defining the behavior within each region) identifications. Based on the approach that these two tasks are carried out, methods to extract rules are further divided into two groups consisting of joint methods and disjoint methods. Joint methods, such as GA (Lawrence 1991), simultaneously identifies the antecedent and consequent by exceeding the capabilities of most optimization algorithms as they can afford the capability of finding global optimal solutions by mimicking biological evolution. As for disjoint methods, the divide and conquer approach is used as the strategy for optimizing the following two tasks: separating and identifying advantages over joint

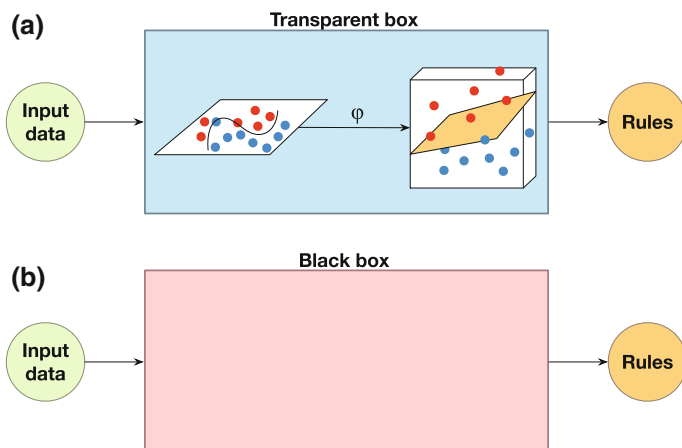


Fig. 8 System flowchart of decompositional and pedagogical rule extraction techniques

ones in computational efficiency. There are three methods that are widely used for partition namely grid (e.g. Wang-Mendel (WM) method (Wang and Mendel 1992)), tree partition (e.g. C4.5 (Quinlan 1993), classification and regression trees (CART), logistic model tree (LMT) and random forest (RF) (Breiman 2001)) as well as clustering (e.g. Mountain clustering and its extension, subtractive clustering (Yager and Filev 1994)).

The interpretability of ANNs and SVMs can be obtained by extracting symbolic rules from the trained model. The rule extraction techniques are used to open up the black box approach by generating symbolic, comprehensible descriptions while maintaining the same predictive power (Martens et al. 2007). Andrews et al. (Andrews 1974; Andrews et al. 1995) proposed an approach for the rule extraction from ANN that can be easily extended to SVMs. Two approaches exist to extract rules from the black-box ANN and SVM models (Martens et al. 2007) which are the decompositional and pedagogical approaches. The decompositional approach determines rules by utilizing information from the internal components of the constructed SVM model while the pedagogical approach considers SVM model as a black box and derives its rules by relating the inputs with the outputs of the SVM model. The difference between the decompositional and pedagogical rule extraction techniques is schematically illustrated in Fig. 8.

For the *decompositional approach*, Setiono and Liu (1995) firstly proposed an approach to understand the ANN's results. Understanding the ANN's results through rule extraction was obtained via the use of a three-phase algorithm as follows: (i), a weight-decay back-propagation network is built such that important connections are reflected by the larger weight values; (ii) the network is pruned by deleting non-informative connections while still maintaining its predictive accuracy; (iii) rules are extracted and produced. In 1997, the decompositional technique NeuroLinear (Setiono and Liu 1997) was developed to extract oblique classification

rules from neural networks comprising of one hidden layer. Kim and Lee (2000) have proposed an algorithm for feature extraction and feature combination by utilizing multilayer perceptron networks with sigmoid functions. A few years later, Gupta et al. (1999) had proposed an analytical framework for classifying existing rule extraction methods for FF-ANN. This method extracts rules by directly interpreting the strengths of the connection weights in a trained network. In the case of the decompositional method, a few research have been published for extracting rules from SVMs. For instance, Núñez et al. (2002) proposed the SVM+Prototypes method for extracting rules from SVMs. The basic idea of this approach consists of: (i) determining the decision function by means of SVM while a clustering algorithm is used to determine prototype vectors for each class; (ii) defining regions in the input space that can be transferred to if-then rules. In 2007, Barakat and Bradley (2007) proposed a novel algorithm for the rule extraction from SVMs known as SQReX-SVM. After training the SVM model, SQReX-SVM directly extracts rules from the support vectors (SVs) by using a modified sequential covering algorithm. Rules are then produced by using the rank of the most discriminative features as measured by the interclass separation.

For the *pedagogical approach*, there are a large number of studies focused on opening the black box nature of ANN as to improve their interpretability. In 1988, Saito and Nakano (1988) have proposed a workflow for medical diagnosis using rule extraction from a modified ANN. A few years later, the BRAINNE system was proposed (Sestito and Dillon 1992) for extracting rules from ANN using back-propagation algorithm. The major contribution of the BRAINNE system is that it can directly deal with continuous data as inputs without requiring discretization. Shortly afterwards, Thrun (1993) proposed the VIA method for extracting rules by mapping inputs directly to the output through the use of a generate-and-test procedure for extracting symbolic rules from ANN trained by the backpropagation algorithm. Furthermore, details on how to improve the interpretability of the black box ANN have been discussed previously (Zhou and Chen 2002; Andrews et al. 1995; Augasta and Kathirvalavakumar 2012). Similar to the case of the decompositional method, only a few studies have been reported for improving the interpretability of ANN via the pedagogical approach. For example, Trepan (Craven and Shavlik 1996) was the first to introduce the pedagogical tree extraction algorithm by extracting decision trees from trained neural networks having an arbitrary architecture. In constructing a tree, this method makes use of the best first expansion strategy to build a tree via recursive partitioning. Trepan allowed splits with at least M-of-N type of tests. At each step, a queue of leaves is further expanded into sub-trees until a stopping criterion is met. In 2007, Martens et al. (2007) proposed the use of an SVM model as an oracle to generate rules. For the convenience of the vast majority of scientists, a MATLAB toolbox for generating rules using any black box model as oracle has been implemented and made publicly available. Previously, many researchers reported that ANN and SVM rule extraction approaches had equal or higher performance when compared with the original ANN and SVM methods (Barakat and Bradley 2007; Augasta and Kathirvalavakumar 2012; Gong et al. 2008).

8.2 White Box Learning Methods

8.2.1 Multiple Linear Regression

MLR is one of the most basic method for performing regression in QSAR modeling. Given a matrix X of a compound of interest, the MLR model assumes that the expected value of Y could be expressed in the form of a linear equation as summarized below:

$$y_i = \sum_{i=1}^m \mathbf{b}_i x_i + \mathbf{b}_0 \quad (12)$$

Generally, this approach is favored for its simplicity and ease of interpretation as the model assumes that there exists a linear relationship between a set of molecular descriptors and the bioactivity. When using MLR, regression coefficients can be obtained via the use of the least squares method. The size of the coefficient may reveal the degree of influence that molecular descriptors has on the bioactivity. Moreover, a positive coefficient indicate that the respective molecular descriptors contributes positively to the bioactivity and vice versa for the negative coefficient. However, in the presence of collinear descriptors, these interpretations may be error prone. A general rule of thumb states that the sample size (i.e. number of compounds in the data set) should be at least five times the number of descriptors that are used.

8.2.2 Logistic Regression

The transformation of MLR to a logistic regression (LR), can be easily performed by representing the Y variable via the conditional probability of Y given X variables ($\pi(X)$) when the logistic distribution is used (Hosmer et al. 2013). The specific formula of LR is defined as follows:

$$\pi(X) = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \dots + b_M x_M}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \dots + b_M x_M}} \quad (13)$$

where \mathbf{b}_i represents the transformation of $\pi(X)$. Furthermore, the logit transformation is defined in terms of $\pi(X)$:

$$\begin{aligned} g(X) &= \ln \left[\frac{\pi(X)}{1 - \pi(X)} \right] \\ &= b_0 + b_1 x_1 + b_2 x_2 + \dots + b_M x_M \end{aligned} \quad (14)$$

For the MLR method, the least square approach is used to estimate unknown parameters \mathbf{b}_i . The basic idea of this method is to minimize the sum of square error between predicted Y and actual Y values. Unfortunately, the least square approach cannot be used to optimize \mathbf{b}_i on a data having a dichotomous variable (i.e. variables

that have a value of 0 or 1). As for the LR method, the maximum likelihood estimator is used to alleviate the problem of dichotomous variables. A convenient way to represent the likelihood probability function for (\mathbf{x}, \mathbf{y}) where $\mathbf{x} = (x_1, x_2, \dots, x_M)$ and $\mathbf{y} = (y_1, y_2, \dots, y_M)$ can be defined as follows:

$$\pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \quad (15)$$

Since the data set (X, Y) is assumed to be independent variables, the likelihood probability function is used to estimate β_i in expressions summarized as follows:

$$l(b_i) = \prod_{i=1}^M \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \quad (16)$$

In the binomial case, where the outputs of LR is close to 0 and 1, respectively, indicates low and high probability of occurrences.

8.2.3 Efficient Linear Method

Efficient linear method (ELM) is a general-purpose learning method proposed by Shoombuatong et al. (2015) that can be used for performing both classification and regression tasks. This approach was first applied in the QSAR study of the bioactivity of aromatase inhibitors (AIs) where it has been shown to afford an interpretable model in which significant features are transparent and can be used to provide insights pertaining to the origin of its bioactivity. The main procedures of the ELM method entails the following steps:

Step 1: Prepare a training data set D_{Tr}^M consisting of positive and negative samples.

Step 2: Formulate a predictive model with a weighted summation $f(C)$ in the form of a linear model as follows:

$$f(C) = \sum_{i=1}^m \mathbf{b}_i x_i + \mathbf{b}_0 \quad (17)$$

Step 3: Select informative features using the fitness function of the Akaike information criterion (AIC). Finally, features affording high feature usage is selected for the construction of a predictive model.

Step 4: Estimate the optimal parameter \mathbf{b} by using the genetic algorithms (GA) with the Andrews' sine function $fitness(x)$ (Andrews 1974). To obtain a reliable parameter, the fitness function utilizes a 10-fold cross-validation (10-fold CV) scheme.

Step 5: Predict the unknown P with the scoring function ($Pred(C)$) using the weighted summation and subsequently discriminate it using only the threshold as obtained from:

$$Pred(C) = \begin{cases} \textit{positive}, & f(C) > \textit{threshold} \\ \textit{negative}, & \textit{otherwise} \end{cases} \quad (18)$$

8.2.4 Principal Component Analysis

The aforementioned learning approach are supervised (e.g. MLR, ANN, or SVM) in which the SAR is discerned from a list of compounds in the training set using the function in the form of $Y = f(X)$ (i.e. Y can be computed as a function of X descriptors). As a counterpart, unsupervised learning methods aim to characterize the underlying patterns of X variables without the need for Y variable. Principal component analysis (PCA) is one of the most commonly used unsupervised learning method for multivariate data analysis that can help reveal details from the high-dimensional information hidden inside the array of numerical descriptors (Jolliffe 2002). PCA analyzes the high-dimensional and intercorrelated X variables and compresses its information into a few dimensions without much loss of the core information while filtering out the noise. Briefly, the first principal component (PC) lies along the direction of maximal data variance capturing the most variability of all possible linear combinations. Because PCA seeks the linear combination of X variables that are uncorrelated with maximal variability, the assumption can be made that the first PC contains the most core information while much of the last PCs contain the noise. PCA focuses on identifying the data structures based on measurement scales and the resulting PC weights will be larger for X variables with higher variation. Two of the most useful features of PCA are the loadings and scores values.

8.2.5 Partial Least Squares Regression

Partial least squares regression (PLSR) is a commonly used learning method for the analysis of large data sets owing to its inherent ability to handle large redundant features and readily produce interpretable regression coefficients from the predictive model. In PCA, only the X variables are considered in the multivariate analysis as it does not take into account the biological properties of compounds (i.e. the Y variable). However, PLSR makes use of the information of Y variables to maximize inter-class variance (Helland 1988). PLSR is a widely used method for constructing predictive models in which features are compressed into orthogonal latent variable or PCs. The origins of PLSR can be traced back to the non-linear iterative partial least squares (NIPALS) algorithm as proposed by Herman Wold (Helland 2001). For the principle assumption of PLSR methods, a data set with intercorrelated variables is generated and then the latent structure are projected by means of PLSR. This learning method can be used for both regression and classification tasks where dimension reduction of the original feature space is an integral part of its modeling process.

8.2.6 Decision Tree

Decision trees (DT) are tree-like graphs that model a decision, which are commonly learned by recursively splitting the set of training instances into subsets based on the instances' values for the explanatory variables (Quinlan 1993). It uses the conditional statement consisting of if-then statement, which allows us to make a prediction. In short, DT constitutes a series of split points that are known as nodes. To make a prediction, we start at the top-most root node, which represents the most important feature. From this root node, a decision threshold value leads to divergence of two subsequent nodes in which the value of the feature of interest is greater than or less than the threshold value. This process is repeated at each subsequent inner nodes until we reach one of the terminal leaf nodes, which are the prediction class (i.e. whether the compound's bioactivity is classified as either being active or inactive).

8.2.7 Random Forest

Random Forest (RF) is an ensemble of unpruned classification and regression tree (Breiman et al. 1984; Breiman 2001). RF takes advantage of two efficient machine learning methods (e.g. bagging and random feature selection). RF is a further development of bagging. Instead of using all features, RF randomly selects two-third of a training data set to build the predictor and the other one-third of the training data set, known as the out-of-bag (OOB) data set, is utilized to evaluate the performance of the predictor. Predictions are derived from the majority vote or averaging the output of all trees for classification and regression problems, respectively. To evaluate the importance for each feature f_i , the values of features f_i in the OOB data set are randomly permuted and the feature importance for f_i can then be evaluated by measuring the decrease of prediction performance of the permuted OOB data set. The prediction performance can be measured by using accuracy or Gini index. The Gini index is calculated by using the impurity of each feature that is capable of separating samples of two (or more) classes. The size of the feature subsets used is a fixed number in which the number of different features tried at each split (m_{try}) are set at $p^{1/2}$ and $p/3$ for classification and regression problems.

9 Resources and Software for Performing QSAR Modeling

In this section, we present some of the software that can be used for the construction of QSAR models. This spans molecular descriptor software, multivariate analysis software and integrated software that typically lowers the steep learning curve that are usually required to get up and running in developing QSAR models.

Prior to the construction of QSAR models, the molecular features of compounds can be discerned via the use of software for computing molecular descriptors. Table 4

Table 4 List of open source descriptor calculation software

	0D	1D	2D	3D	PCP	Availability	Ref.
CDK	✓	✓	✓	✓	✓	Java, R and Python	Guha (2017), rcdk (2017), O'Boyle and Hutchison (2008)
RCPI	✓	✓	✓	✓	✓	R	Xiao et al. (2017)
ChemmineR	✓	✓	✓	✓	✓	R	Girke (2017)
PaDEL	✓	✓	✓	✓	✓	Java, Standalone	Yap (2017), Yap (2011)
ChemDes	✓	✓	✓	✓	✓	Web server	Cao (2017a), Dong et al. (2015)
jCompoundMapper		✓	✓	✓		Java	Hinselmann et al. (2017), Hinselmann et al. (2011)
QuBiL _s -MAS			✓			Standalone	Ponce (2017a), Medina Marrero et al. (2015)
QuBiL _s -MIDAS				✓		Standalone	Ponce (2017b), Garcia-Jacas et al. (2014)
Chemical Descriptors Library (CDL)	✓	✓	✓	✓	✓	C++ library	Molplex Ltd. and Sykora (2017)
ChemoPy	✓	✓	✓	✓	✓	Web server	Cao (2017b), Cao et al. (2013)
Pybel	✓	✓	✓	✓	✓	Python	O'Boyle et al. (2017a), Oldham et al. (2008)
Babel	✓	✓	✓	✓	✓	Standalone	O'Boyle et al. (2017b, 2011)

Table 5 Summary of software for performing QSAR modeling

Software	Description	Standalone	Online	Ref.
AutoWeka	Automated data mining software based on Weka machine learning package	✓		Nantasenamat et al. (2015)
AZOrange	Open source high performance machine learning in a graphical environment	✓		Stalring et al. (2011)
CDK-Taverna	Platform independent workflow environment for cheminformatics	✓		Kuhn et al. (2010)
CHARMMing	Aside from ligand docking this suite of tools supports QSAR model building		✓	Miller et al. (2008)
ChemBench	Web platform for building QSAR models		✓	Walker et al. (2010)
ChemMine	Cheminformatics and data mining tools for small molecule data analysis		✓	Backman et al. (2011)
CORAL	Software for building QSAR models using SMILES-based descriptors via Monte Carlo	✓		Benfenati et al. (2011)
DMax Chemistry Assistant	Data mining tool for QSAR, compound data analysis and virtual screening	✓		DTAI Research Group (2017)
MOE Cheminformatics and QSAR	Module for performing cheminformatics and QSAR modeling	✓		Chemical Computing Group Inc. (2017)
OCHEM	Online platform for building QSAR models	✓	✓	Sushko et al. (2011)
OCED QSAR Toolbox	QSAR application toolbox for assessing hazards of chemicals	✓		Dimitrov et al. (2016)
PASS Online	Predicts the biological activity spectra of query compounds	✓	✓	Filimonov et al. (2014)
QSARINS	QSAR modeling tool in agreement with OECD principles	✓		Gramatica et al. (2013)
QSAR Workbench	QSAR workflow tool with numerical and graphical results	✓		Cox et al. (2013)
Toxtree	Toxicity estimation using decision tree	✓		Patlewicz et al. (2008)

Table 6 Summary of software for multivariate analysis

Software	Description	License	Ref.
Benchmarkware	Data mining software for analyzing biological and chemical data	Commercial	Certara (2017)
ChemmineR	Cheminformatics package for analyzing drug-like small molecule data in R	Free	Cao et al. (2008)
IBM SPSS	Statistical and data mining software for multivariate data analysis	Commercial	IBM (2017)
KEEL	Java-based software for performing various	Free	Alcal-Fdez et al. (2011)
KNIME	Modular data exploration and mining platform that allow users to create data flows and extend functionality via modular API	Free	Mazanetz et al. (2012)
LIBSVM	Data mining software based on SVM algorithm	Free	Chang and Lin (2011)
Neuralware	Platform for developing and deploying empirical modeling based on neural networks	Commercial	NeuralWare (2017)
Neural Network Toolbox	MATLAB package providing algorithms, functions, and tools to create, train, visualize and simulate neural networks	Commercial	The MathWorks, Inc. (2017a)
MAPLE	Mathematical and computational engine with an intuitive user interface	Commercial	Maplesoft (2017)
MATLAB	Interactive environment and programming language for performing computationally intensive tasks visual programming on Python scripting	Commercial	The MathWorks, Inc. (2017b)
PyChem	Python package for chemometric for univariate and multivariate data analysis	Free	Jarvis et al. (2006)
R	Comprehensive statistical environment for data analysis and graphics visualization	Free	Ripley (2017)
RapidMiner	Open source system for data mining with an intuitive graphical user interface	Free	RapidMiner, Inc. (2017)

(continued)

Table 6 (continued)

Software	Description	License	Ref.
SAS Enterprise Miner	Reveal insights from data mining analysis	Commercial	SAS Institute Inc. (2017)
Scikit-learn	Python package for data mining analysis	Free	Pedregosa et al (2017)
SNNS	Software simulator for neural networks on Unix workstations	Free	Zell et al. (2017)
SOM Toolbox	MATLAB package for implementation the self-organizing map algorithm and more	Free	Kohonen (2017)
Spotfire S+	Statistical programming environment for analysis large scale data as well as an interactive graphics system for creation of statistical charts	Commercial	TIBCO Software Inc. (2017)
The Unscrambler	Chemometric software for data analysis and design of experiments	Commercial	CAMO Software AS (2017)
WEKA	Java-based software for data analysis via a wide range of machine learning algorithm	Free	Frank et al. (2017)

Table 7 Comparison of machine learning packages and modules from R, Python's scikit-learn and WEKA

Methods	R package	Python's scikit-learn	Weka
SVM	e1071	SVC, NuSVC and LinearSVC	LibSVM
ANN	neuralnet	MLPClassifier and MLPRegressor	MultilayerPerceptron
DL	deeplearning	–	–
MLR	car	LinearRegression	LinearRegression
LR	logistf	LogisticRegression	Logistic
ELM	<i>R script</i> ^a	–	–
PCA	princomp	PCA	PrincipalComponents
PLSR	pls	PLSRegression	PartialLeastSquares
DT	C50	DecisionTreeClassifier and DecisionTreeRegressor	J48graft
RF	randomForest	RandomForestClassifier and RandomForestRegressor	RandomForest

^a<http://dx.doi.org/10.6084/m9.figshare.1274030>

summarizes the available software along with the dimensional type of descriptor that can be computed.

A wide range of software and tools for performing QSAR modeling are available as either standalone desktop-based application or as web-based application as summarized in Table 5.

Table 6 lists some of the software for performing multivariate analysis for computer savvy scientists as the software may require a steeper learning curve than those listed in Table 5.

Table 7 summarizes the comparison between three popular machine learning packages in three popular languages namely R, Python and Java.

10 Conclusion

In spite of certain inherent flaws, the QSAR paradigms inevitably is one of the driving forces contributing to the advancements in drug discovery and design. As with all technologies, QSAR is not perfect, however, its weaknesses and flaws are continuously being identified, solved and reformed to help shape a more robust QSAR model. Particularly, the present chapter argues for the increased use of interpretable QSAR models in drug discovery research. QSAR models were originally intended to assist medicinal chemists with design ideas that are often overlooked as a useful approach; one reason is that chemists and biologists do not understand the underlying assumptions of the predictions. Hence, we have presented several concepts pertaining to inverse QSAR techniques that can reconstruct a chemical

structure with good synthetic feasibility based on features identified by QSAR models. We have also presented concepts on rule extraction methods that can unravel the black box and make interpretations of machine learning approaches. Furthermore, we reviewed the utility of various molecular descriptors in the post-genomic era of the biological data deluge. Moreover, the concept of conformal prediction have also been discussed as a novel and potentially powerful approach that can define the relative confidence or reliability of predictions made. The inherent heterogeneity and vagueness of details describing the construction of QSAR models in the literature may hinder further progress. Therefore, markup language such as QSAR-ML have been suggested as a means to solve the reproducibility of QSAR models by standardizing and demystifying the underlying details of QSAR models (i.e. addition of metadata on the source of the data set, the type of descriptors used, the machine learning employed, software names and version that are used, etc.) as well as making them exchangeable (i.e. in the context that they can be shared and readily be used by the scientific community). The availability of interpretable molecular descriptors and transparent machine learning methods presents a positive outlook for the utility of QSARs in drug discovery research. The application of several key sets of standards in QSAR modeling will further help to enhance their generalization and acceptance by the wider drug research community.

Acknowledgements This work is supported by a Research Career Development Grant (No. RSA5780031) to CN from the Thailand Research Fund; the New Scholar Research Grant (No. MRG5980220) to WS from the Thailand Research Fund; and the Swedish Research Links program (No. C0610701) to CN and JESW from the Swedish Research Council.

References

- Alcal-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., Garca, S., Snchez, L., et al. (2011). *Journal of Multiple-Valued Logic and Soft Computing*, 17, 255.
- Andrews, D. F. (1974). *Technometrics*, 16(4), 523.
- Andrews, R., Diederich, J., & Tickle, A. B. (1995). *Knowledge-Based Systems*, 8(6), 373.
- Augasta, M. G., & Kathirvalavakumar, T. (2012). *Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, Salem, Tamilnadu* (pp. 21–23).
- Backman, T. W., Cao, Y., & Girke, T. (2011). *Nucleic Acids Research*, 39, W486.
- Baell, J. B., & Holloway, G. A. (2010). *Journal of Medicinal Chemistry*, 53(7), 2719.
- Bajorath, J. (2014). *Molecular Informatics*, 33(6–7), 438.
- Balaban, A. T. (1982). *Chemical Physics Letters*, 89(5), 399.
- Barakat, N. H., & Bradley, A. P. (2007). *IEEE Transactions on Knowledge and Data Engineering*, 19(6), 729.
- Barakat, N., & Bradley, A. P. (2010). *Neurocomputing*, 74(1), 178.
- Barakat, N., & Diederich, J. (2004). *14th International Conference on Computer Theory and Applications (ICCTA'2004)*. Alexandria, Egypt.
- Barakat, N., & Diederich, J. (2005). *International Journal of Computational Intelligence*, 2(1), 59.
- Benfenati, E., Toropov, A. A., Toropova, A. P., Manganaro, A., & Gonella, D. R. (2011). *Chemical Biology and Drug Design*, 77(6), 471.
- Bengio, Y. (2009). *Foundations and Trends in Machine Learning*, 2(1), 1.
- Borman, S. (1990). *Chemical and Engineering News*, 68(8), 20.

- Bordes, A., Chopra, S., & Weston, J. (2014). arXiv preprint: [arXiv:1406.3676](https://arxiv.org/abs/1406.3676).
- Bravi, G., Gancia, E., Mascagni, P., Pegna, M., Todeschini, R., & Zaliani, A. (1997). *Journal of Computer-Aided Molecular Design*, 11(1), 79.
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. New York, USA: CRC Press.
- Breiman, L. (2001). *Machine Learning*, 45(1), 5.
- Broto, P., Moreau, G., & Vanduycke, C. (1984). *European Journal of Medicinal Chemistry*, 19(1), 66.
- Brown, N., McKay, B., & Gasteiger, J. (2006). *Journal of Computer-Aided Molecular Design*, 20(5), 333.
- Brustle, M., Beck, B., Schindler, T., King, W., Mitchell, T., & Clark, T. (2002). *Journal of Medicinal Chemistry*, 45(16), 3345.
- Burges, C. J. (1998). *Data Mining and Knowledge Discovery*, 2(2), 121.
- Cao, D. S. (2017a). ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation. <http://www.scbdd.com/chemdes>.
- Cao, D. S. (2017b). ChemoPy Descriptor Calculator. http://www.scbdd.com/chemopy_desc/index/.
- Cao, Y., Charisi, A., Cheng, L. C., Jiang, T., & Girke, T. (2008). *Bioinformatics*, 24(15), 1733.
- Cao, D., Liang, Y., Xu, Q., Yun, Y., & Li, H. (2011). *Journal of Computer-Aided Molecular Design*, 25(1), 67.
- Cao, D. S., Xu, Q. S., Hu, Q. N., & Liang, Y. Z. (2013). *Bioinformatics*, 29(8), 1092.
- CAMO Software AS. (2017). The Unscrambler. <http://www.camo.com/rt/Products/Unscrambler/unscrambler.html>.
- Capuzzi, S. J., Politi, R., Isayev, O., Farag, S., & Tropsha, A. (2016). *Frontiers of Environmental Science*, 4, 3.
- Certara. (2017). Benchware 3D Explorer. <https://www.certara.com/software/molecular-modeling-and-simulation/benchware-3d-explorer/>.
- Chang, C. C., & Lin, C. J. (2011). *ACM Transactions on Intelligent Systems and Technology*, 2(27), 1.
- Charoenkwan, P., Shoombuatong, W., Lee, H. C., Chaijaruwanich, J., Huang, H. L., & Ho, S. Y. (2013). *PLoS One*, 8(9), e72368.
- Chemical Computing Group Inc. (2017). Molecular Operating Environment (MOE). https://www.chemcomp.com/MOE-ChemInformatics_and_QSAR.htm.
- Chen, H., Carlsson, L., Eriksson, M., Varkonyi, P., Norinder, U., & Nilsson, I. (2013). *Journal of Chemical Information and Modeling*, 53(6), 1324.
- Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., et al. (2014). *Journal of Medicinal Chemistry*, 57(12), 4977.
- Chiu, Y. Y., Lin, C. T., Huang, J. W., Hsu, K. C., Tseng, J. H., You, S. R., et al. (2013). *Nucleic Acids Research*, 41(Database issue), D430.
- Churchwell, C. J., Rintoul, M. D., Martin, S., Visco, D. P., Kotu, A., Larson, R. S., et al. (2004). *Journal of Molecular Graphics and Modelling*, 22(4), 263.
- Ciodaro, T., Deva, D., De Seixas, J., & Damazio, D. (2012). *Journal of Physics: Conference Series*, 368, 012030. IOP Publishing.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). *Journal of Machine Learning Research*, 12, 2493.
- Consonni, V., Todeschini, R., & Pavan, M. (2002). *Journal of Chemical Information and Computer Sciences*, 42(3), 682.
- Cortes-Ciriano, I., Ain, Q. U., Subramanian, V., Lenselink, E. B., Mendez-Lucio, O., IJzerman, A. P., et al. (2015). *Medicinal Chemical Communications*, 6, 24.
- Cortes, C., & Vapnik, V. (1995). *Machine Learning*, 20(3), 273.
- Costello, J. C., Heiser, L. M., Georgii, E., Gonen, M., Menden, M. P., Wang, N. J., et al. (2014). *Nature Biotechnology*, 32(12), 1202.
- Cox, R., Green, D. V., Luscombe, C. N., Malcolm, N., & Pickett, S. D. (2013). *Journal of Computer-Aided Molecular Design*, 27(4), 321.

- Cramer, R. D., Patterson, D. E., & Bunce, J. D. (1988). *Journal of the American Chemical Society*, 110(18), 5959.
- Craven, M. W., & Shavlik, J. W. (1996). *Advances in neural information processing systems* (pp. 24–30). Cambridge, USA: MIT Press.
- Cros, A. F. A. (1863). Action de l'alcohol amylique sur l'organisme. Ph.D. thesis, University of Strasbourg.
- Crum-Brown, A., & Fraser, T. (1868). *Transactions of the Royal Society of Edinburgh*, 25, 151.
- Danishuddin, A. U. K. (2016). *Drug Discovery Today*, 21(8), 1291.
- Dearden, J., Cronin, M., & Kaiser, K. (2009). *SAR and QSAR in Environmental Research*, 20(3–4), 241.
- de Vries, S. J., van Dijk, M., & Bonvin, A. M. (2010). *Nature Protocols*, 5(5), 883.
- Destrero, A., Mosci, S., De Mol, C., Verri, A., & Odone, F. (2009). *Computational Management Science*, 6(1), 25.
- Devinyak, O., Havrylyuk, D., & Lesyk, R. (2014). *Journal of Computer-Aided Molecular Design*, 54, 194.
- Dimova, D., & Bajorath, J. (2016). *Molecular Informatics*, 35(5), 181.
- Dimitrov, S. D., Didericj, R., Sobanski, T., Pavlov, T. S., Chapkov, G. V., Chapkonov, A. S., et al. (2016). *SAR and QSAR in Environmental Research*, 1–17.
- Dong, J., Cao, D. S., Miao, H. Y., Liu, S., Deng, B. C., Yun, Y. H., et al. (2015). *Journal of Cheminformatics*, 7, 60.
- Doweyko, A. M. (2004). *Journal of Computer-Aided Molecular Design*, 18(7), 587.
- Doweyko, A. M. (2008). *Journal of Computer-Aided Molecular Design*, 22(2), 81.
- Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., Vapnik, V. (1996). *Proceedings of the 9th International Conference on Neural Information Processing Systems, NIPS'96* (pp. 155–161). Cambridge, MA, USA: MIT Press.
- DTAI Research Group (2017). DMax Chemistry Assistant. <https://dtai.cs.kuleuven.be/software/dmax/>.
- Durant, J. L., Leland, B. A., Henry, D. R., & Nourse, J. G. (2002). *Journal of Chemical Information and Computer Sciences*, 42(6), 1273.
- Ebrahimi, E., Monjezi, M., Khalesi, M. R., & Armaghani, D. J. (2016). *Bulletin of Engineering Geology and the Environment*, 75(1), 27.
- Eklund, M., Norinder, U., Boyer, S., & Carlsson, L. (2012). In L. Iliadis, I. Maglogiannis, H. Papadopoulos, K. Karatzas, & S. Sioutas (Eds.), *Artificial Intelligence Applications and Innovations: AIAI 2012 International Workshops: AIAB, AIEA, CISE, COPA, IIVC, ISQL, MHDW, and WADTMB, Halkidiki, Greece, September 27–30, 2012, Proceedings, Part II* (pp. 166–175). Berlin, Germany: Springer.
- Eklund, M., Norinder, U., Boyer, S., & Carlsson, L. (2014). *Journal of Chemical Information and Modeling*, 54(3), 837.
- Eriksson, M., Chen, H., Carlsson, L., Nissink, J. W., Cumming, J. G., & Nilsson, I. (2014). *Journal of Chemical Information and Modeling*, 54(4), 1117.
- Esbensen, K. H., & Geladi, P. (2010). *Journal of Chemometrics*, 24(3–4), 168.
- Faulon, J. L. (1994). *Journal of Chemical Information and Computer Sciences*, 34(5), 1204.
- Faulon, J. L. (1996). *Journal of Chemical Information and Computer Sciences*, 36(4), 731.
- Faulon, J. L., Churchwell, C. J., & Visco, D. P. (2003). *Journal of Chemical Information and Computer Sciences*, 43(3), 721.
- Faulon, J. L., Collins, M. J., & Carr, R. D. (2004). *Journal of Chemical Information and Computer Sciences*, 44(2), 427.
- Faulon, J. L., Brown, W. M., & Martin, S. (2005). *Journal of Computer-Aided Molecular Design*, 19(9–10), 637.
- Feng, B. Y., Shelat, A., Doman, T. N., Guy, R. K., & Shoichet, B. K. (2005). *Nature Chemical Biology*, 1(3), 146.
- Feng, B. Y., Simeonov, A., Jadhav, A., Babaoğlu, K., Inglese, J., Shoichet, B. K., et al. (2007). *Journal of Medicinal Chemistry*, 50(10), 2385.

- Feng, B. Y., & Shoichet, B. K. (2006). *Nature Protocols*, 1(2), 550.
- Filimonov, D. A., Zakharov, A. V., Lagunin, A. A., & Poroikov, V. V. (2009). *SAR and QSAR in Environmental Research*, 20(7), 679.
- Filimonov, D. A., Lagunin, A. A., Glorizova, T. A., Rudik, A. V., Druzhilovskii, D. S., Pogodin, P. V., et al. (2014). *Chemistry of Heterocyclic Compounds*, 50(3), 444.
- Frank, E., Hall, M. & Trigg, L. Weka. <http://www.cs.waikato.ac.nz/ml/weka/>.
- Free, S. M., & Wilson, J. W. (1964). *Journal of Medicinal Chemistry*, 7(4), 395.
- Fu, X., Ong, C., Keerthi, S., Hung, G. G., & Goh, L. (2004). In *Proceedings of IEEE International Joint Conference on Neural Networks* (pp. 291–296). Budapest, Hungary: IEEE.
- Fujita, T., & Winkler, D. A. (2016). *Journal of Chemical Information and Modeling*, 56(2), 269.
- Fung, G., Sandilya, S., & Rao, R. B. (2005). *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (pp. 32–40). New York, USA: ACM.
- Gallup, G. A., Gilkerson, W., & Jones, M. (1952). *Transactions of the Kansas Academy of Science*, 55(2), 232.
- Gao, H., Katzenellenbogen, J. A., Garg, R., & Hansch, C. (1999). *Chemical Reviews*, 99(3), 723.
- Garcia-Jacas, C. R., Marrero-Ponce, Y., Acevedo-Martinez, L., Barigye, S. J., Valdes-Martini, J. R., & Contreras-Torres, E. (2014). *Journal of Computational Chemistry*, 35(18), 1395.
- Garg, R., Gupta, S. P., Gao, H., Babu, M. S., Debnath, A. K., & Hansch, C. (1999). *Chemical Reviews*, 99(12), 3525.
- Garg, R., Kurup, A., Mekapati, S. B., & Hansch, C. (2003). *Chemical Reviews*, 103(3), 703.
- Geronikaki, A. A., Lagunin, A. A., Hadjipavlou-Litina, D. I., Eleftheriou, P. T., Filimonov, D. A., Poroikov, V. V., et al. (2008). *Journal of Medicinal Chemistry*, 51(6), 1601.
- Girke, T. (2017). ChemmineR: Cheminformatics toolkit for R. <https://www.bioconductor.org/packages/release/bioc/html/ChemmineR.html>.
- Gleeson, M. P. (2008). *Journal of Medicinal Chemistry*, 51(4), 817.
- Gobbi, M., Beeg, M., Toropova, M. A., Toropov, A. A., & Salmons, M. (2016). *Toxicology Letters*, 250, 42.
- Golbraikh, A., Fourches, D., Sedykh, A., Muratov, E., Liepina, I., & Tropsha, A. (2014). *Practical aspects of computational chemistry III* (pp. 187–230). Boston, USA: Springer.
- Gong, R., Huang, S. H., & Chen, T. (2008). *IEEE Transactions on Industrial Informatics*, 4(3), 198.
- Gonzalez, M. P., Tern, C., Fall, Y., Teijeira, M., & Besada, P. (2005). *Bioorganic and Medicinal Chemistry*, 13(3), 601.
- Goodarzi, M., Heyden, Y. V., & Funar-Timofei, S. (2013). *Trends in Analytical Chemistry*, 42, 49.
- Gozalbes, R., Doucet, J. P., & Derouin, F. (2002). *Current Drug Targets Infectious Disorders*, 2(1), 93.
- Gramatica, P., Chirico, N., Papa, E., Cassani, S., & Kovarich, S. (2013). *Journal of Computational Chemistry*, 34(24), 2121.
- Guha, R. (2017). CDK Descriptor Calculator GUI (version 1.4. 6). <http://www.rguha.net/code/java/cdkdesc.html>.
- Guha, R., & Van Drie, J. H. (2008). *Journal of Chemical Information and Modeling*, 48(8), 1716.
- Gupta, A., Park, S., & Lam, S. M. (1999). *IEEE Transactions on Knowledge and Data Engineering*, 11(6), 985.
- Güttele, M., Helma, C., Karwath, A., & Kramer, S. (2013). *Molecular Informatics*, 32(5–6), 516.
- Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). *Machine Learning*, 46(1–3), 389.
- Guyon, I. (2003). *Journal of Machine Learning Research*, 3, 1157.
- Hadjipavlou-Litina, D., Garg, R., & Hansch, C. (2004). *Chemical Reviews*, 104(9), 3751.
- Hall, M. A. (1999). Correlation-based feature selection for machine learning. Ph.D. thesis, The University of Waikato.
- Hammitt, L. P. (1937). *Journal of the American Chemical Society*, 59(1), 96.
- Hansch, C., Maloney, P. P., Fujita, T., & Muir, R. M. (1962). *Nature*, 194, 178.
- Hansch, C., Leo, A., & Taft, R. (1991). *Chemical Reviews*, 91(2), 165.
- Hansch, C., Hoekman, D., & Gao, H. (1996). *Chemical Reviews*, 96(3), 1045.

- Hansch, C., Hoekman, D., Leo, A., Weininger, D., & Selassie, C. D. (2002). *Chemical Reviews*, 102(3), 783.
- Hansch, C. (2011). *Journal of Computer-Aided Molecular Design*, 25(6), 495.
- Hansch, C., & Gao, H. (1997). *Chemical Reviews*, 97(8), 2995.
- Harding, A. P., Wedge, D. C., & Popelier, P. L. (2009). *Journal of Chemical Information and Modeling*, 49(8), 1914.
- Hawkins, D. M., Basak, S. C., & Mills, D. (2003). *Journal of Chemical Information and Computer Sciences*, 43(2), 579.
- Héberger, K., & Rajkó, R. (2002). *Journal of Chemometrics*, 16(8), 436.
- Heberger, K., & Skrbic, B. (2012). *Analytica Chimica Acta*, 716, 92.
- Helland, I. S. (1988). *Communication in Statistics: Simulation and Computation*, 17(2), 581.
- Helland, I. S. (2001). *Chemometrics and Intelligent Laboratory*, 58(2), 97.
- Helmstaedter, M., Briggman, K. L., Turaga, S. C., Jain, V., Seung, H. S., & Denk, W. (2013). *Nature*, 500(7461), 168.
- Hinselmann, G., Rosenbaum, L., Jahn, A., Fechner, N., & Zell, A. (2011). *Journal of Cheminformatics*, 3(1), 3.
- Hinselmann, G., Rosenbaum, L., Jahn, A., Fechner, N., & Zell, A. (2017). jCompoundMapper: An open source java library and command-line tool for chemical fingerprints. <http://jcompoundmapper.sourceforge.net/>.
- Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). *Neural Computing*, 18(7), 1527.
- Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. (2012). arXiv preprint [arXiv:1207.0580](https://arxiv.org/abs/1207.0580).
- Hosmer, D. W, Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (pp. 1–33). New Jersey, USA: Wiley.
- Hu, X., Hu, Y., Vogt, M., Stumpfe, D., & Bajorath, J. (2012). *Journal of Chemical Information and Modeling*, 52(5), 1138.
- IBM. (2017). IBM SPSS Software. <http://www.ibm.com/analytics/us/en/technology/spss/>.
- Jarvis, R. M., Broadhurst, D., Johnson, H., O'Boyle, N. M., & Goodacre, R. (2006). *Bioinformatics*, 22(20), 2565.
- Jelfs, S., Ertl, P., & Selzer, P. (2007). *Journal of Chemical Information and Modeling*, 47(2), 450.
- Johnson, S. R. (2008). *Journal of Chemical Information and Modeling*, 48(1), 25.
- Jolliffe, I. (2002). *Principal component analysis*. New York, USA: Springer.
- Katritzky, A. R., Kuanar, M., Slavov, S., Hall, C. D., Karelson, M., Kahn, I., et al. (2010). *Chemical Reviews*, 110(10), 5714.
- Khan, M. T., & Sylte, I. (2007). *Current Drug Discovery Technologies*, 4(3), 141.
- Kier, L. B., & Hall, L. H. (1976). *Molecular connectivity in chemistry and drug research*. New York, USA: Academic Press.
- Kim, K. H. (2007a). *Journal of Computer-Aided Molecular Design*, 21(8), 421.
- Kim, K. H. (2007b). *Journal of Computer-Aided Molecular Design*, 21(1–3), 63.
- Kim, D., & Lee, J. (2000). In López de Mántaras and Plaza (Eds.), *Proceedings of the 11th European conference on machine learning* (pp. 211–219). London, UK: Springer.
- Kohonen, T. (2017). SOM: Self-Organization Map. <http://www.cis.hut.fi/somtoolbox/>.
- Krasavin, M. (2015). *European Journal of Medicinal Chemistry*, 97, 525.
- Kubinyi, H. (1988). *Quantitative Structure-Activity Relationship*, 7(3), 121.
- Kubinyi, H. (1993). *3D QSAR in drug design: Volume 1: Theory methods and applications* (Vol. 1). Dordrecht, Netherlands: Springer Science & Business Media.
- Kubinyi, H. (2006). In S. Ekins (Ed.) *Computer applications in pharmaceutical research and development* (pp. 377–424). New Jersey, USA: Wiley.
- Kufareva, I., & Abagyan, R. (2008). *Journal of Medicinal Chemistry*, 51(24), 7921.
- Kuhn, T., Willighagen, E. L., Zielesny, A., & Steinbeck, C. (2010). *BMC Bioinformatics*, 11, 159.
- Kurgan, L., Razib, A. A., Aghakhani, S., Dick, S., Mizianty, M., & Jahandideh, S. (2009). *BMC Structural Biology*, 9, 50.
- Kurup, A., Garg, R., & Hansch, C. (2000). *Chemical Reviews*, 100(3), 909.

- Kurup, A., Garg, R., Carini, D. J., & Hansch, C. (2001). *Chemical Reviews*, 101(9), 2727.
- Kurup, A., Garg, R., & Hansch, C. (2001). *Chemical Reviews*, 101(8), 2573.
- Kvasnicka, V., & Pospichal, J. (1996). *Journal of Chemical Information and Computer Sciences*, 36(3), 516.
- Lawrence, D., et al. (1991). *Handbook of genetic algorithms*. New York, USA: Van No Strand Reinhold.
- Leung, M. K., Xiong, H. Y., Lee, L. J., & Frey, B. J. (2014). *Bioinformatics*, 30(12), i121.
- Li, Q., Wang, Y., & Bryant, S. H. (2009). *Bioinformatics*, 25(24), 3310.
- Lipnick, R. L. (1991). *Studies of narcosis*. Dordrecht, Netherlands: Springer.
- Liu, S. S., Yin, C. S., Li, Z. L., & Cai, S. X. (2001). *Journal of Chemical Information and Computer Sciences*, 41(2), 321.
- Liu, H., & Motoda, H. (2007). *Computational methods of feature selection*. Boca Raton, Florida: CRC Press.
- Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., & Svetnik, V. (2015). *Journal of Chemical Information and Modeling*, 55(2), 263.
- Manallack, D. T. (2008). *Perspectives in Medicinal Chemistry*, 1, 25.
- Maplesoft. (2017). Maple. <https://www.maplesoft.com/products/Maple/>.
- Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. (2007). *European Journal of Operational Research*, 183(3), 1466.
- Masand, V. H., Toropov, A. A., Toropova, A. P., & Mahajan, D. T. (2014). *Current Computer-Aided Drug Design*, 10, 75.
- Mazanetz, M. P., Marmon, R. J., Reisser, C. B., & Morao, I. (2012). *Current Topics in Medicinal Chemistry*, 12(8), 1965.
- McGovern, S. L., Caselli, E., Grigorieff, N., & Shoichet, B. K. (2002). *Journal of Medicinal Chemistry*, 45(8), 1712.
- Medina Marrero, R., Marrero-Ponce, Y., Barigye, S. J., Echeverria Diaz, Y., Acevedo-Barrios, R., Casanola-Martin, G. M., et al. (2015). *SAR and QSAR in Environmental Research*, 26(11), 943.
- Miller, B. T., Singh, R. P., Klauda, J. B., Hodoscek, M., Brooks, B. R., & Woodcock, H. L. (2008). *Journal of Chemical Information and Modeling*, 48(9), 1920.
- Molplex Ltd., & Sykora, V. (2017). Chemical Descriptors Library (CDL). <https://sourceforge.net/projects/cdelib/>.
- Morgenthaler, M., Schweizer, E., Hoffmann-Roder, A., Benini, F., Martin, R. E., Jaeschke, G., et al. (2007). *ChemMedChem*, 2(8), 1100.
- Nantasenamat, C., Naenna, T., Isarankura-Na-Ayudhya, C., & Prachayasittikul, V. (2005). *Journal of Computer-Aided Molecular Design*, 19(7), 509.
- Nantasenamat, C., Isarankura-Na-Ayudhya, C., Naenna, T., & Prachayasittikul, V. (2007a). *Biosensors and Bioelectronics*, 22(12), 3309.
- Nantasenamat, C., Isarankura-Na-Ayudhya, C., Tansila, N., Naenna, T., & Prachayasittikul, V. (2007b). *Journal of Computational Chemistry*, 28(7), 1275.
- Nantasenamat, C., Piacham, T., Tantimongcolwat, T., Naenna, T., Isarankura-Na-Ayudhya, C., & Prachayasittikul, V. (2008). *Journal of Biological Systems*, 16(02), 279.
- Nantasenamat, C., Isarankura-Na-Ayudhya, C., Naenna, T., & Prachayasittikul, V. (2009). *EXCLI Journal*, 8(7), 74.
- Nantasenamat, C., Isarankura-Na-Ayudhya, C., & Prachayasittikul, V. (2010). *Expert Opinion on Drug Discovery*, 5(7), 633.
- Nantasenamat, C., Worachartcheewan, A., Jamsak, S., Preeyanon, L., Shoombuatong, W., Simeon, S., et al. (2015). In H. Cartwright (Ed.), *Artificial neural networks* (pp. 119–147). New York, NY, USA: Springer.
- Nantasenamat, C., & Prachayasittikul, V. (2015). *Expert Opinion on Drug Discovery*, 10(4), 321.
- NeuralWare. (2017). NeuralWare. <http://www.neuralware.com/>.
- Núñez, H., Angulo, C., & Català, A. (2002). *10th European Symposium on Artificial Neural Networks (ESANN)*, pp. 107–112.
- O'Boyle, N. M., & Hutchison, G. R. (2008). *Chemistry Central Journal*, 2, 24.

- O'Boyle, N. M., Morley, C., & Hutchison, G. R. (2008). *Chemistry Central Journal*, 2, 5.
- O'Boyle, N. M., Morley, C. & Hutchison, G. R. (2017a). Pybel. https://openbabel.org/docs/dev/UseTheLibrary/Python_Pybel.html.
- O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). *Journal of Cheminformatics*, 3, 33.
- O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2017b). Open Babel: The open source chemistry toolbox. <http://openbabel.org/>.
- Oldham, M. C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., et al. (2008). *Nature Neuroscience*, 11(11), 1271.
- Patani, G. A., & LaVoie, E. J. (1996). *Chemical Reviews*, 96(8), 3147.
- Patlewicz, G., Jeliaskova, N., Safford, R. J., Worth, A. P., & Aleksiev, B. (2008). *SAR and QSAR in Environmental Research*, 19(5–6), 495.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M. et al. (2017). Scikit-learn. <http://scikit-learn.org/>.
- Peng, H., Long, F., & Ding, C. (2005). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), 1226.
- Poroikov, V. V., Filimonov, D. A., Ihlenfeldt, W. D., Glorizova, T. A., Lagunin, A. A., Borodina, Y. V., et al. (2003). *Journal of Chemical Information and Computer Sciences*, 43(1), 228.
- Prachayasittikul, V., Worachartcheewan, A., Shoombuatong, W., Songtawe, N., Simeon, S., Prachayasittikul, V., et al. (2015). *Current Topics in Medicinal Chemistry*, 15(18), 1780.
- Prathipati, P., Pandey, G., & Saxena, A. K. (2005). *Journal of Chemical Information and Modeling*, 45(1), 136.
- Prathipati, P., Dixit, A., & Saxena, A. K. (2007). *Journal of Computer-Aided Molecular Design*, 92, 29.
- Prathipati, P., Ma, N. L., & Keller, T. H. (2008). *Journal of Chemical Information and Modeling*, 48(12), 2362.
- Prathipati, P., & Mizuguchi, K. (2016a). *Current Topics in Medicinal Chemistry*, 16(9), 1009.
- Prathipati, P., & Mizuguchi, K. (2016b). *Journal of Chemical Information and Modeling*, 56(6), 974.
- Prathipati, P., & Saxena, A. K. (2005). *Journal of Computer-Aided Molecular Design*, 19(2), 93.
- Pudil, P., Novovičová, J., & Kittler, J. (1994). *Pattern Recognition Letters*, 15(11), 1119.
- Ponce, Y. M. (2017a). QuBiLS-MAS. <http://tomocomd.com/qubils-mas>.
- Ponce, Y. M. (2017b). QuBiLS-MIDAS. <http://tomocomd.com/qubils-midas>.
- Qiu, T., Qiu, J., Feng, J., Wu, D., Yang, Y., Tang, K., et al. (2016). *Briefings in Bioinformatics*, 18(1), 125.
- Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, USA: Morgan Kaufmann Publishers Inc.
- RapidMiner, Inc. (2017). RapidMiner. <https://rapidminer.com/>.
- rdck: Interface to the CDK Libraries. <https://cran.r-project.org/web/packages/rdck/index.html>.
- Rácz, A., Bajusz, D., & Héberger, K. (2015). *SAR and QSAR in Environmental Research*, 26(7–9), 683.
- Radoux, C. J., Olsson, T. S., Pitt, W. R., Groom, C. R., & Blundell, T. L. (2016). *Journal of Medicinal Chemistry*, 59(9), 4314.
- Raiko, T., Valpola, H., & LeCun, Y. (2012). In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*. *JMLR Workshop and Conference Proceedings* (Vol. 22, pp. 924–932).
- Randic, M. (1975). *Journal of the American Chemical Society*, 97(23), 6609.
- Ripley, B. D. (2017). The R project in statistical computing. <https://www.stats.ox.ac.uk/pub/bdr/LTSN-R.pdf>.
- Rogers, D., & Hahn, M. (2010). *Journal of Chemical Information and Modeling*, 50(5), 742.
- Rosenbaum, L., Hinselmann, G., Jahn, A., & Zell, A. (2011). *Journal of Cheminformatics*, 3(1), 11.

- Rucker, C., Rucker, G., & Meringer, M. (2007). *Journal of Chemical Information and Modeling*, 47(6), 2345.
- Rueda, M., Bottegoni, G., & Abagyan, R. (2009). *Journal of Chemical Information and Modeling*, 49(3), 716.
- Rueda, M., Bottegoni, G., & Abagyan, R. (2010). *Journal of Chemical Information and Modeling*, 50(1), 186.
- Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., & Todeschini, R. (2012). *Molecules*, 17(5), 4791.
- Sahigara, F., Ballabio, D., Todeschini, R., & Consonni, V. (2013). *Journal of Cheminformatics*, 5(1), 27.
- Saito, K., & Nakano, R. (1988). In *IEEE International Conference on Neural Networks, 1988* (pp. 255–262). IEEE.
- SAS Institute Inc. (2017). SAS Enterprise Miner. http://www.sas.com/en_th/software/analytics/enterprise-miner.html.
- Saxena, A. K., & Prathipati, P. (2003). *SAR and QSAR in Environmental Research*, 14(5–6), 433.
- Saxena, A. K., & Prathipati, P. (2006). *SAR and QSAR in Environmental Research*, 17(4), 371.
- Schuffenhauer, A., Brown, N., Selzer, P., Ertl, P., & Jacoby, E. (2006). *Journal of Chemical Information and Modeling*, 46(2), 525.
- Seebeck, B., Wagener, M., & Rarey, M. (2011). *ChemMedChem*, 6(9), 1630.
- Selassie, C. D., Garg, R., Kapur, S., Kurup, A., Verma, R. P., Mekapati, S. B., et al. (2002). *Chemical Reviews*, 102(7), 2585.
- Sestito, S., & Dillon, T. (1992). *Proceedings of the 12th International Conference on Expert Systems and their Applications (AVIGNON'92)* (pp. 645–656).
- Setiono, R., & Liu, H. (1995). *Proceedings of the 14th International Joint Conference on Artificial Intelligence—Volume 1, IJCAI'95* (pp. 480–485). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Setiono, R., & Liu, H. (1997). *Neurocomputing*, 17(1), 1.
- Setiono, R., Leow, W. K., & Zurada, J. M. (2002). *IEEE Transactions on Neural Networks*, 13(3), 564.
- Shafer, G., & Vovk, V. (2008). *Journal of Machine Learning Research*, 9, 371.
- Sheridan, R. P. (2015). *Journal of Chemical Information and Modeling*, 55(6), 1098.
- Sheridan, R. P., & Kearsley, S. K. (1995). *Journal of Chemical Information and Computer Sciences*, 35(2), 310.
- Shoombuatong, W., Prachayasittikul, V., Prachayasittikul, V., & Nantasenamat, C. (2015). *EXCLI Journal*, 14, 452.
- Shoombuatong, W., Prachayasittikul, V., Anuwongcharoen, N., Songtawee, N., Monnor, T., Prachayasittikul, S., et al. (2015). *Drug Design. Development and Therapy*, 9, 4515.
- Siedlecki, W., & Sklansky, J. (1988). *International Journal of Pattern Recognition and Artificial Intelligence*, 2(02), 197.
- Simeon, S., Möller, R., Almgren, D., Li, H., Phanus-umporn, C., Prachayasittikul, V., et al. (2016a). *Chemometrics and Intelligent Laboratory Systems*, 151, 51.
- Simeon, S., Spjuth, O., Lapins, M., Nabu, S., Anuwongcharoen, N., Prachayasittikul, V., et al. (2016b). *PeerJ*, 4, e1979.
- Simpson, P. K. (1990). *Artificial neural system: Foundation, paradigm, application and implementations*. Pennsylvania, USA: Windcrest/McGraw-Hill.
- Sippl, W. (2006). *Molecular interaction fields* (pp. 145–170). KGaA: Wiley-VCH Verlag GmbH & Co.
- Skvortsova, M. I., Baskin, I. I., Slovokhotova, O. L., Palyulin, V. A., & Zefirov, N. S. (1993). *Journal of Chemical Information and Computer Sciences*, 33(4), 630.
- Sliwoski, G., Mendenhall, J., & Meiler, J. (2016). *Journal of Computer-Aided Molecular Design*, 30(3), 209.
- Song, M., Breneman, C. M., Bi, J., Sukumar, N., Bennett, K. P., Cramer, S., et al. (2002). *Journal of Chemical Information and Computer Sciences*, 42(6), 1347.

- Spjuth, O., Willighagen, E. L., Guha, R., Eklund, M., & Wikberg, J. E. (2010). *Journal of Cheminformatics*, 2, 5.
- Spyrakakis, F., & Cavasotto, C. N. (2015). *Archives of Biochemistry and Biophysics*, 583, 105.
- Stalring, J. C., Carlsson, L. A., Almeida, P., & Boyer, S. (2011). *Journal of Cheminformatics*, 3, 28.
- Standfuss, J., Edwards, P. C., D'Antona, A., Fransen, M., Xie, G., Oprian, D. D., et al. (2011). *Nature*, 471(7340), 656.
- Stumpfe, D., Hu, Y., Dimova, D., & Bajorath, J. (2014). *Journal of Medicinal Chemistry*, 57(1), 18.
- Sushko, I., Novotarskyi, S., Krner, R., Pandey, A. K., Rupp, M., et al. (2011). *Journal of Computer-Aided Molecular Design*, 25(6), 533.
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K., & Q. Weinberger (Ed.) *Advances in neural information processing systems* 27 (pp. 3104–3112). Curran Associates, Inc.
- The MathWorks, Inc. (2017a). Neural Network Toolbox. <http://www.mathworks.com/products/neural-network/>.
- The MathWorks, Inc. (2017b). MATLAB. <https://www.mathworks.com/products/matlab/>.
- TIBCO Software Inc. (2017). TIBCO Spotfire S+. <http://spotfire.tibco.com/discover-spotfire/who-uses-spotfire/by-role/statisticians>.
- Tarca, A. L., Than, N. G., & Romero, R. (2013). *Systems Biomedicine*, 1(4), 217.
- Taskinen, J., & Yliruusi, J. (2003). *Advanced Drug Delivery Reviews*, 55(9), 1163.
- Thornber, C. W. (1979). *Chemical Society Reviews*, 8(4), 563.
- Thorne, N., Auld, D. S., & Inglese, J. (2010). *Current Opinion in Chemical Biology*, 14(3), 315.
- Thrun, S. (1993). *Extracting provably correct rules from artificial neural networks*. Bonn, Germany: University of Bonn.
- Todeschini, R., & Consonni, V. (2008). *Handbook of molecular descriptors*. Weinheim, Germany: Wiley-VCH Verlag GmbH.
- Toropov, A. A., Toropova, A. P., Benfenati, E., Leszczynska, D., & Leszczynski, J. (2010). *Journal of Computational Chemistry*, 31(2), 381.
- Toropov, A. A., Toropova, A. P., Puzyn, T., Benfenati, E., Gini, G., Leszczynska, D., et al. (2013). *Chemosphere*, 92(1), 31.
- Toropova, A. P., & Toropov, A. A. (2014). *European Journal of Pharmaceutical Sciences*, 52, 21.
- Toropov, A. A., & Benfenati, E. (2007a). *European Journal of Medicinal Chemistry*, 42(5), 606.
- Toropov, A. A., & Benfenati, E. (2007b). *Current Drug Discovery Technologies*, 4(2), 77.
- Tosco, P., Balle, T., & Shiri, F. (2011). *Journal of Computer-Aided Molecular Design*, 25(8), 777.
- Tropsha, A., Gramatica, P., & Gombar, V. K. (2003). *QSAR and Combinatorial Science*, 22(1), 69.
- Tropsha, A. (2010). *Molecular Informatics*, 29(6–7), 476.
- Venkatasubramanian, V., Chan, K., & Caruthers, J. M. (1995). *Journal of Chemical Information and Computer Sciences*, 35(2), 188.
- Verma, R. P., & Hansch, C. (2005). *Bioorganic and Medicinal Chemistry*, 13(15), 4597.
- Verma, R. P., & Hansch, C. (2009). *Chemical Reviews*, 109(1), 213.
- Visco, D. P., Pophale, R. S., Rintoul, M. D., & Faulon, J. L. (2002). *Journal of Molecular Graphics and Modelling*, 20(6), 429.
- Walker, T., Grulke, C. M., Pozefsky, D., & Tropsha, A. (2010). *Bioinformatics*, 26(23), 3000.
- Wang, L. X., & Mendel, J. M. (1992). *IEEE Transactions on Systems, Man and Cybernetics: Systems*, 22(6), 1414.
- Wei, D. B., Zhang, A. Q., Han, S. K., & Wang, L. S. (2001). *SAR and QSAR in Environmental Research*, 12(5), 471.
- Weis, D. C., Faulon, J. L., LeBorne, R. C., & Visco, D. P. (2005). *Industrial and Engineering Chemistry*, 44(23), 8883.
- Wong, W. W., & Burkowski, F. J. (2009). *Journal of Cheminformatics*, 1, 4.
- Worachartcheewan, A., Nantasenamat, C., Naenna, T., Isarankura-Na-Ayudhya, C., & Prachayasitikul, V. (2009). *European Journal of Medicinal Chemistry*, 44(4), 1664.

- Worachartcheewan, A., Mandi, P., Prachayasittikul, V., Toropova, A. P., Toropov, A. A., & Nantasenamat, C. (2014). *Chemometrics and Intelligent Laboratory Systems*, 138, 120.
- Worachartcheewan, A., Prachayasittikul, V., Toropova, A. P., Toropov, A. A., & Nantasenamat, C. (2015). *Molecular Diversity*, 19(4), 955.
- Worth, A. P., & Cronin, M. T. (2004). *Alternatives to Laboratory Animals*, 32, 703.
- Xiao, N., Cao D. S., & Xu, Q. (2017). Rcp: Toolkit for compound-protein interaction in drug discovery. <http://bioconductor.org/packages/release/bioc/html/Rcpi.html>.
- Xing, L., Glen, R. C., & Clark, R. D. (2003). *Journal of Chemical Information and Computer Sciences*, 43(3), 870.
- Yager, R. R., & Filev, D. P. (1994). *Journal of Intelligent & Fuzzy Systems*, 2(3), 209.
- Yap, C. W. (2011). *Journal of Computational Chemistry*, 32(7), 1466.
- Yap, C. W. (2017). PaDEL-Descriptor. <http://www.yapcsoft.com/dd/padeldescriptor>.
- Zakharov, A. V., Peach, M. L., Sitzmann, M., & Nicklaus, M. C. (2014). *Journal of Chemical Information and Modeling*, 54(3), 705.
- Zell, A., Mache, N., Hubner, R., Mamier, G., Vogt, M., Döring, S., et al. (2017). SNNS: Stuttgart neural network simulator. <http://www.ra.cs.uni-tuebingen.de/SNNS/>.
- Zhao, Z., Wu, H., Wang, L., Liu, Y., Knapp, S., Liu, Q., et al. (2014). *ACS Chemical Biology*, 9(6), 1230.
- Zhang, Y., Su, H., Jia, T., & Chu, J. (2005). In Ho T. B., Cheung D., Liu H. (Eds.), *Advances in knowledge discovery and data mining: 9th Pacific-Asia conference on knowledge discovery and data mining* (pp. 61–70). Berlin/Heidelberg, Germany: Springer.
- Zhou, Z. H., & Chen, S. F. (2002). *Journal of Research and Development*, 39(4), 398.
- Zurada, J. M. (1992). *Introduction to artificial neural systems* (Vol. 8). Minnesota, USA: West Publishing Co.

The Use of Topological Indices in QSAR and QSPR Modeling

John C. Dearden

Abstract Topological indices (TIs) are numerical representations of the topology of a molecule, and are calculated from the heavy atom graphical depiction of the molecule. One of the first TIs was that of Wiener in 1947, who showed that his index correlated well with the boiling points of alkanes. There are now many different TIs available, and many of them are discussed in this chapter, with respect largely to their use as descriptors in QSAR/QSPR modeling. Three types in particular stand out, molecular connectivities developed by Randić and Kier and Hall, electrotopological state (E-state) values developed by Kier and Hall, and information content indices developed by Basak and co-workers. New TIs are still appearing, despite some criticism that there are already too many types of TI, that they are difficult of interpretation, and that they are inferior to physicochemical descriptors in modeling.

Keywords Wiener • Randić • Information content • Molecular connectivity • Electrotopological state • Biodescriptors • Inverse QSAR • Software • Hostility to topological indices

1 Introduction

1.1 What Is QSAR?

QSARs (quantitative structure-activity relationships) and QSPRs (quantitative structure-property relationships) are mathematical correlations between a specified biological activity or molecular property and one or more physicochemical and/or

J.C. Dearden (✉)

School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University,
Byrom Street, Liverpool L3 3AF, UK
e-mail: j.c.dearden@ljmu.ac.uk

© Springer International Publishing AG 2017

K. Roy (ed.), *Advances in QSAR Modeling*, Challenges and Advances
in Computational Chemistry and Physics 24, DOI 10.1007/978-3-319-56850-8_2

molecular structural properties, known as descriptors since they “describe” the activity or property under examination. A simple example is given by Eq. 1 (Cronin et al. 2002) for the inhibition of growth of the aquatic ciliate *Tetrahymena pyriformis* by phenols:

$$\begin{aligned} \log(1/IGC_{50}) &= 0.53 \log D - 0.96 LUMO - 0.58 \\ n = 160 \quad r^2 &= 0.81 \quad q^2 = 0.80 \quad s = 0.34 \quad F = 340 \end{aligned} \quad (1)$$

where IGC_{50} = concentration of a substituted phenol required to inhibit growth by 50%, D = its distribution coefficient between 1-octanol and water buffered to pH 7.35, $LUMO$ = energy of the lowest unoccupied molecular orbital of the chemical (a measure of electrophilicity), n = number of chemicals used to develop the QSAR (termed the training set), r = correlation coefficient, q = cross-validated correlation coefficient, s = standard error of prediction by the QSAR, and F = Fisher statistic (variance ratio).

For those not familiar with QSAR, a brief explanation of Eq. 1 and its accompanying statistics is apposite. Activities and toxicities are almost invariably used in QSAR as the logarithm of the reciprocal concentration (or dose) to produce a required effect, for two reasons: (a) activities and toxicities can range over many orders of magnitude, so taking logarithms makes the numbers easier to handle, and (b) QSARs can be considered as modifications of the van't Hoff isotherm, which relates the free energy change in a process to the logarithm of the equilibrium or rate constant controlling the process; QSPRs in particular are sometimes referred to as linear free energy relationships (LFERs) (Wells 1968).

The statistical information provided gives an indication of the goodness-of-fit, robustness and predictive ability of the QSAR model. The coefficient of determination, r^2 , is a measure of how well the QSAR models the data; a value of 0.81 means that the model explains 81% of the variation in $\log IGC_{50}$. Considering that IGC_{50} is a measure of in vivo toxicity, that is a very good value; one only rarely finds r^2 values much greater than 0.8 when modeling in vivo data, because of inherent error in the data. q^2 is an internal cross-validated coefficient of determination, an indicator of how predictive the QSAR is—that is, how well it predicts IGC_{50} values for chemicals that were not used to develop the QSAR. This is done by removing one chemical from the training set, re-developing the QSAR without that chemical, and then using the re-developed QSAR to predict the IGC_{50} value of the removed chemical. That chemical is then returned to the training set, another chemical is removed and the process repeated until all chemicals have been removed in turn. A combined predictive indicator q^2 is then calculated. It should be noted that q^2 is not now considered to be a good indicator of predictivity (Golbraikh and Tropsha 2002). It is preferable to use chemicals that have not been used at all to develop a QSAR model, in order to test its predictivity.

Some years ago a set of principles, the OECD Principles for the Validation of (Q)SARs, was devised as guidance for the development and use of QSARs (OECD 2004):

A valid QSAR/QSPR should have:

- (1) *a defined endpoint;*
- (2) *an unambiguous algorithm;*
- (3) *a defined domain of applicability;*
- (4) *appropriate measures of goodness-of-fit, robustness and predictivity;*
- (5) *a mechanistic interpretation, if possible.*

The OECD report went on to say:

It is recognised that it is not always possible, from a scientific viewpoint, to provide a mechanistic interpretation of a given (Q)SAR (Principle 5)...The absence of a mechanistic interpretation for a model does not mean that a model is not potentially useful in the regulatory context. The intent of Principle 5 is not to reject models that have no apparent mechanistic basis, but to ensure that some consideration is given to the possibility of a mechanistic association between the descriptors used in a model and the endpoint being predicted, and to ensure that this association is documented.

It is important to note that “it is not always possible...to provide a mechanistic interpretation of a given (Q)SAR”. There are far too many examples in the literature of descriptors being wrongly interpreted in an attempt to provide a mechanistic interpretation. Johnson (2008) commented that “QSAR has devolved into a perfectly practiced art of logical fallacy: *cum hoc ergo propter hoc* (with this, therefore because of this)”.

1.2 What Are Topological Indices?

There are now thousands of physicochemical and structural descriptors available for use in QSAR/QSPR modeling. The vast majority of these are calculated values, since experimental measurement is time-consuming and expensive, whereas calculation is rapid and less expensive with the wide range of software now available for that purpose (see Table 1). Among these calculated values are the descriptors known as topological indices, which are graph invariants that encode the topology of molecules depicted as graphs (Devillers 1999a), usually without hydrogen atoms (i.e., hydrogen-suppressed graphs). In such graphs, atoms are termed ‘vertices’ and bonds are termed ‘edges’. The graphs are two-dimensional (2D), as shown in Fig. 1, and show the non-hydrogen atoms and their connections with each other (their connectivity).

In order to calculate a TI, typically the values of adjacent vertices, or some function of them such as square root or reciprocal, are multiplied, and then summed across all edges. So, for the 2-methylpentane molecule shown in Fig. 1, a simple TI would be $1 \times 3 + 1 \times 3 + 3 \times 2 + 2 \times 2 + 2 \times 1$, which is 18.

Many different types of topological indices are now available, through the manipulation of adjacency matrices and distance matrices (across multiple edges), including the use of graphs with more than one edge between at least one pair of

Table 1 Some software for calculation of topological indices

Software	Indices calculated ^a	Website (all accessed on 11 July 2016)
ADAPT	χ (0–7) ^b , kappa (1–3), E-state, Wiener	http://research.chem.psu.edu/pcjgroup/adapt.html
ADMET Predictor	E-state	http://www.simulations-plus.com
ADMEWORKS Predictor	χ (0–7), kappa (1–3), E-state, Wiener	http://www.fqs.pl
Bluedesc	χ , WHIM, autocorrelation	http://www.ra.cs.uni-tuebingen.de/software/bluedesc/welcome_e.html
ChemDes	χ , kappa, E-state, information content, WHIM, autocorrelation	http://www.scbdd.com/chemdes/
Chemistry Development Kit	χ (0–1), kappa, WHIM	http://www.opentox.org/dev/documentation/components/cdk
ChemProp	E-state	http://www.ufz.de/index.php?en=34593
CODESSA	χ , kappa, flexibility, Wiener, Balaban J, information content	http://www.semichem.com
CORINA Symphony	Autocorrelation	http://www.mn.am.com/products/corinasymphony
Dragon	χ , E-state, Randić, Zagreb, information content, ETA, autocorrelation	http://www.talete.mi.it/
JOELib	E-state, kappa, autocorrelation, Zagreb	http://www.ra.cs.uni-tuebingen.de/software/joelib/index.html
MathChem	χ , Zagreb, Randić, Balaban J, Wiener	https://pypi.python.org/pypi/mathchem
MDL QSAR	χ (0–10), kappa (1–3), flexibility, Shannon, Wiener, Platt	https://www.mdl.com/products/predictive/qsar/index.jsp
Molconn-Z	χ (0–10), kappa (1–3), flexibility, Shannon, Wiener, Platt	https://www.edusoft-lc.com/molconn/
Mold ²	χ , flexibility, Zagreb, Randić, Balaban J, Wiener, autocorrelation, information content	https://www.fda.gov/ScienceResearch/BioinformaticsTools/Mold2
Molecular Modeling Pro	χ (0–4), kappa (2), E-state, Wiener	https://www.chemistry-software.com
MOE (Molecular Operating Environment)	χ (0–1), kappa (0–3), flexibility, E-state, Wiener, Balaban J, Zagreb	http://www.chemcomp.com
MOLE db	χ , E-state, Randić, Zagreb, information, ETA, autocorrelation, WHIM	http://michem.disat.unimib.it/mole_db/
PaDEL-Descriptor	χ , kappa (1–3), E-state, Wiener, Zagreb, WHIM	http://padel.nus.edu.sg/software/padeldescriptor

(continued)

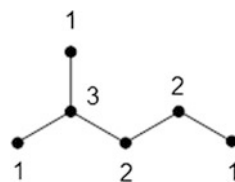
Table 1 (continued)

Software	Indices calculated ^a	Website (all accessed on 11 July 2016)
POLLY	information content (0–6), χ (0–6), Wiener	No website; copyright of University of Minnesota, 1988
PreADMET	χ , kappa, Wiener, Balaban J	https://preadmet.bmdrc.kr/
QSARPro	χ , kappa, information content	http://www.vlifesciences.com/products/QSARPro/Product_QSARpro.php
QuaSAR	χ (1, 2), kappa (1–3), flexibility, Wiener, Zagreb, Balaban J	http://www.chemcomp.com/journal/descr.htm
RDKit	χ (0–4), kappa (1–3), Balaban J	https://rdkit.readthedocs.io/en/latest/
SciQSAR	χ , kappa, E-state	http://www.pharmaceuticalonline.com/doc/sciqsar-2d-0001
T.E.S.T.	χ (0–10), kappa (1–3), E-state, information content, autocorrelation, Zagreb, Balaban J, Wiener	http://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test
TOPIX	χ (0–8), kappa (1–3), Randić, Wiener, Zagreb	http://www.lohninger.com/topix.html
VCCLAB Parameter Client	χ , kappa (1–3), flexibility, E-state, information content, Wiener, Randić, Balaban J, Zagreb, centric, autocorrelation, WHIM	http://www.vcclab.org/lab/pclient

^aSome software will calculate other topological descriptors in addition to those listed

^b χ (0–7) means χ values for path lengths 0 to 7; χ means no path lengths specified

Fig. 1 Hydrogen-suppressed graph of 2-methylpentane, with labelled vertices



adjacent vertices (i.e. taking account of double and triple bonds), termed multi-graphs, and weighted graphs, whereby the contributions of various edges and/or vertices are modified according to the relative importance of each to the TI. Many have proved useful in QSAR modeling (Netzeva 2004). An early example is the modeling of the potency of non-specific local anaesthetics by Kier et al. (1975), using an approach developed by Randić (1975). They obtained an excellent correlation of minimum blocking concentrations with what is now termed simple first order molecular connectivity, χ :

$$\begin{aligned}\log MBC &= 3.55 - 0.762\chi \\ n = 36 \quad r^2 &= 0.966 \quad s = 0.390\end{aligned}\tag{2}$$

The calculation of χ values is described in Sect. 16.

This chapter considers the usefulness of the main topological indices employed in QSAR and QSPR modeling. An early review of the subject is that of Balaban (1985). Those readers interested in the philosophy and theory of topological indices should consult Ivanciuc and Balaban (1999), Basak (2013a) and Roy et al. (2015). Rouvray and King (2002) and Todeschini and Consonni (2009) have presented and discussed, *inter alia*, a very wide range of topological indices. It should be noted that TIs are real numbers that represent aspects of molecular structure (Yilmaz and Götürk 2009), and thus are qualified for use as descriptors in QSAR/QSPR modeling.

1.3 The Value of Topological Indices

Topological indices have been shown to correlate well with numerous biological and physicochemical properties, suggesting that they are information-rich, and they are also generally quickly and readily calculated. They are therefore useful descriptors in QSARs and QSPRs that are used for predictive purposes, such as prediction of the toxicity of a chemical or the potency of a drug for future release in the market.

However, the term “descriptor” can be taken to relate not only to statistical description of the dependent variable, but also to physicochemical and/or structural description, implying that the descriptor(s) can yield information about the process (es) that control the magnitude of the dependent variable.

In connection with topological indices, Kubinyi (1993) forcefully pointed out that “in contrast to general recommendations on the selection of biologically meaningful parameters (descriptors), the physicochemical meaning of the topological parameters is never clear”. He later (Kubinyi 1997) described them as having a “hidden secret”. Livingstone (2000) has pointed out that with such large numbers of topological indices available for QSAR/QSPR use, there is a danger of chance correlations occurring. Lopez de Compadre et al. (1983) pointed out that there are dangers in their application to non-homologous series.

It has to be acknowledged that the inability of topological descriptors to allow much if any physicochemical interpretation is a grave drawback. Nevertheless, as Devillers (1999a) has pointed out, “these problems do not (mean) that topological indices must not be used in QSAR and QSPR studies. Indeed, they only show that, like all the other molecular descriptors, they have to be employed only in contexts for which they are suitable”. So long as this is recognised and acted upon, the use of topological indices is valid and valuable. Indeed, Randić et al. (2016) have made a

powerful case for the use of topological indices as true indicators of molecular structure in QSAR and QSPR modeling.

Of course, some TIs have proved more useful than others. In the present author's view, three types have proved especially valuable, namely (i) information content indices, developed by Basak and co-workers (see Sect. 7); (ii) molecular connectivities, devised initially by Randić as a branching index (see Sect. 13), and developed by Kier and Hall (see Sect. 14); and (iii) electrotopological state (E-state) indices, developed by Kier and Hall (see Sect. 19). The main types of TI in use today are discussed below in semi-chronological order of their introduction.

2 The Wiener Index

One of the first topological indices (Ivanciuc 2000) to be used in QSPR modeling is the Wiener index W (Wiener 1947), which gives an additive measure of the connections in a hydrogen-suppressed molecular graph. It is defined for hydrocarbons in terms of two variables, (a) the polarity number p , which is the number of pairs of carbon atoms that are separated by three carbon-carbon bonds, and (b) the path number w , calculated as follows: multiply the total number of carbon atoms on one side of any bond by those on the other side, and sum these for all bonds.

With this approach Wiener was able to predict the boiling points of a series of branched and straight chain paraffins to within an average of 1° , using Eq. 3:

$$\Delta t = (98/n^2) \Delta w + 5.5 \Delta p \quad (3)$$

where Δt = difference in boiling point between a straight and a branched chain isomer, and w and p = structural variables. Wiener later (Wiener 1948) used the same type of equation to model other physicochemical properties of isomeric alkanes. For example, for surface tension he found an average error of prediction of $0.13 \text{ dyne.cm}^{-1}$. The Wiener index has also been correlated with critical constants (Stiel and Thodos 1962), density and viscosity (Rouvray and Crafford 1976), and van der Waals surface area (Gutman and Körtvélyesi 1995).

Ivanciuc (2000) used two Wiener descriptors, weighted to account for the presence of heteroatoms and multiple bonds, along with $\log D$, to model the toxicity of 47 nitrobenzenes to *T. pyriformis*, with $r^2 = 0.875$ and $s = 0.250$, which is comparable to the model developed by Dearden et al. (1995) using physicochemical descriptors ($\log D$, $LUMO$ and modulus of change of charge on the nitro oxygen atom upon substitution), with $r^2 = 0.867$ and $s = 0.255$. Dearden et al. (1995) were able to make mechanistic interpretations of their results, concluding that the nitrobenzenes were behaving as pro-electrophiles, whilst Ivanciuc (2000) was unable to do so, as the Wiener descriptors yield little or no mechanistic information.

3 The Platt and Gordon-Scantlebury Indices

Platt (1947) devised a simple scheme to predict the physicochemical properties of alkanes, by summing the number of adjacent bonds for each atom. The index is not widely used, although Bharate and Singh (2011) found that it contributed to several good QSAR models of the anti-leishmanial effect of phloroglucinol-terpene adducts.

The Gordon-Scantlebury index (1964) is defined as the number of distinct ways that a sequence of three bonds can be overlapped on to the carbon skeleton of a molecule. Sabljic (1990) has pointed out that the index is equal to half the value of the Platt index. Like the Platt index, the Gordon-Scantlebury index is little used in QSAR. One instance of its use is an investigation of the antimycobacterial activity of alkenols (Gupta et al. 2005), although it did not compare well with other topological indices in that work.

4 The Hosoya Index

The Hosoya index Z is the number of sets of non-adjacent bonds in a molecule (Hosoya 1971). In other words, the Hosoya index of a hydrogen-suppressed graph is the total number of matchings within the graph, where a matching is a subset of edges that do not share a vertex. Like other topological indices, it gives a measure of molecular branching. Solomon et al. (2009) used it, along with a number of other descriptors, to model the inhibition of cholinesterase activity, although their best models included the Wiener index rather than the Hosoya index.

5 The Zagreb Indices

The first Zagreb index is calculated simply as the sum of the squares of the number of non-hydrogen bonds formed by each heavy atom (Gutman and Trinajstić 1972). A number of modifications of this were later developed, and Nikolić et al. (2003) have discussed these in detail. Whilst there has been much discussion on the derivation of Zagreb indices, there are relatively few publications that demonstrate their value in QSAR/QSPR modeling (Singh et al. 2014). Two such are that of Bajaj et al. (2005), who used refined Zagreb indices to model the anti-inflammatory activity of *N*-arylanthranilic acids, and that of Dureja et al. (2008), who found that the Zagreb topochemical indices M_1 and M_2 were valuable in modeling the fraction bound and clearance of cephalosporins in humans.

6 The Balaban J Index

The Balaban index (Balaban 1982), also called the average distance sum connectivity index, is computed as follows: the numbers of edges from atom i to all other atoms in a molecule are summed, and this procedure is repeated for all other atoms. The sums for adjacent atoms are then multiplied together, and the reciprocal square roots taken and summed. This number is then multiplied by $(B/(C + 1))$, where B = number of bonds in the molecule and C = number of rings, to give the Balaban index J . Unlike other topological indices, it does not increase rapidly with molecular size (Maran et al. 2010).

Mekenyan et al. (1987) found that J did not correlate well with a range of physicochemical properties or with acute toxicity of ethers. However, Thakur et al. (2004) found that the inhibition of carbonic anhydrase by sulphonamides was modeled well by J :

$$\begin{aligned} \log K_c &= 31.41 - 9.619 J \\ n = 29 \quad r^2 &= 0.910 \quad s = 0.429 \quad F = 274.2 \end{aligned} \quad (4)$$

where K_c = inhibition constant.

7 Information Content Indices

Shannon (1948) was probably the first to study the science of information theory, which relates to molecular complexity (Mowshowitz 1968). However, the main proponents of this approach are undoubtedly Subhash Basak of the University of Minnesota and his co-workers (Basak 1999, 2013b and references cited therein).

There are four main types of information indices: mean, total, complementary and structural information content (Basak 1999). The reader is referred to the publications of Basak (1999, 2013b) for details of the calculation of these indices. Like most topological indices, information content indices appear to work best with homologous series, as with the examples cited by Basak (1987), such as the anesthetic potency (AD_{50}) of barbiturates;

$$\begin{aligned} AD_{50} &= 0.33TIC_1 - 0.002(TIC_1)^2 - 18.50 \\ n = 13 \quad r^2 &= 0.98 \quad s = 0.06 \end{aligned} \quad (5)$$

where TIC_1 = first-order total information content.

This is statistically a slightly better model than that obtained using $\log P$ (octanol-water partition coefficient):

$$\begin{aligned}AD_{50} &= 1.58 \log P - 0.44(\log P)^2 + 1.93 \\ n = 13 \quad r^2 &= 0.94 \quad s = 0.10\end{aligned}\tag{6}$$

A combination of information indices and other descriptors can yield good QSAR models for diverse data sets. For example, for acute toxicity to fathead minnow of 69 diverse benzene derivatives (Gute and Basak 1997), a combination of topostructural and topochemical indices yielded $r^2 = 0.783$ and $s = 0.36$. When geometric and quantum chemical descriptors were added, the correlation improved: $r^2 = 0.863$, $s = 0.30$. However, geometric and quantum chemical descriptors alone did not yield good models.

8 Autocorrelation Descriptors

The autocorrelation approach, first introduced by Moreau and Broto (1980), derives molecular descriptors encoding various physicochemical or structural properties from the molecular graphs of the organic chemicals being studied (Devillers 1999b). The procedure is as follows:

- (1) The shortest interatomic distances, expressed as number of bonds, between each pair of atoms i and j are calculated.
- (2) An appropriate physicochemical or structural property is chosen, and the autocorrelation vector is calculated as the sum of the products of the atomic contributions to that property for each distance between the different atoms.

Many physicochemical properties have been used in QSARs and QSPRs involving this approach (Devillers 1999b). González et al. (2006) used masses, electronegativities and van der Waals volumes to model the inhibitory activity of cytokinin-derived cyclin-dependent kinase inhibitors.

Abreu et al. (2009) used a combination of autocorrelation descriptors and radial distribution descriptors to model the radical scavenging activity of benzo[*b*]thiophenes. In a comparative assessment of 2D autocorrelation, CoMFA and CoMSIA modeling of protein tyrosine kinase inhibition, Caballero et al. (2008) found that CoMSIA performed best.

9 WHIM Descriptors

WHIM (Weighted Holistic Invariant Molecular) descriptors are geometrical descriptors based on statistical indices calculated on the projections of atoms along principal axes (Todeschini et al. 1994). They are able to capture relevant molecular

information regarding molecular size, shape, symmetry and atom distribution with respect to invariant reference frames. In the WHIM approach a molecule is considered as a configuration of points (the atoms) in the three-dimensional space defined by the Cartesian axes. Projections of the atoms along each principal axis are made, and their distributions around the geometric centre are evaluated.

Todeschini and Gramatica (1997) found that WHIM descriptors performed very well in the prediction of a number of physicochemical properties of chlorophenols and of their aquatic toxicity to a range of species. Vlaia et al. (2009) obtained excellent correlations of WHIM descriptors with the toxicity of 48 aliphatic esters to the aquatic ciliate *Tetrahymena pyriformis*. Tong et al. (2008) used the vectors of principal component scores of WHIM indices of peptide analogues to model properties such as bitter taste ($n = 48$, $r^2 = 0.873$, RMSE (root mean square error) = 0.225) and bactericidal activity ($n = 12$, $r^2 = 0.997$, RMSE = 0.133).

10 Topochemical Atom Indices

Topochemically Arrived Unique (TAU) indices, developed by Pal et al. (1988), take account of the chemistry of the atomic core and valence electronic environment of atoms. A detailed explanation of their derivation has been given by Roy and Saha (2003), who used them to model the aqueous solubility of 193 diverse acyclic compounds, with excellent results ($r^2 = 0.946$, $s = 0.735$). Roy and Ghosh (2003) then extended the scope of the TAU scheme by redefining its basic parameters and introducing a novel Extended Topochemical Atom (ETA) formalism. They showed that their new ETA indices could model the toxicities of 50 substituted phenols to *Tetrahymena pyriformis* very well ($r^2 = 0.948$, $q^2 = 0.936$, $s = 0.161$, $F = 159.1$). They also found (Roy and Ghosh 2004) a good correlation of ETA indices with acute toxicity of substituted benzenes to the guppy ($r^2 = 0.885$, $q^2 = 0.865$, $s = 0.230$, $F = 92.6$), and a good correlation (Roy and Ghosh 2009) of a combination of 15 ETA and non-ETA topological descriptors with the toxicity of 288 diverse aromatic compounds to *Tetrahymena pyriformis* ($r^2 = 0.854$, $q^2 = 0.821$, is not given, $F = 106.3$). It may be noted that the use of a large number of descriptors in a QSAR is not recommended (Aptula et al. 2005) as it makes interpretation difficult.

11 The Centric Index

The centric index C developed by Balaban (1979) reflects molecular shape. It uses a procedure known as pruning partition of terminal atoms, and is calculated as follows:

$$C = \Sigma(a_i)^2 \quad (7)$$

where a_i = number of atoms deleted in step i .

The index has not been widely used in QSAR and QSPR investigations. Two studies that employed it (Jalali-Heravi and Asadollahi-Baboli 2008; Noorizadeh et al. 2011) found it not to feature in their best models.

12 Triplet Indices

Filip et al. (1987) introduced a new approach for obtaining graph invariants with very high discrimination ability, called triplet indices. Local vertex invariants (LOVIs) are assembled into a triplet TI, based on one of several operations such as: (i) summation; (ii) summation of squares; (iii) summation of square roots, and so on (Basak et al. 2000). Filip et al. (1987) showed that triplet indices correlated well with physicochemical properties such as boiling points, and NMR chemical shifts.

13 The Randić Index

Randić (1975) devised a topological index to characterize branching in alkanes. In Fig. 2 are depicted the hydrogen-suppressed graphs of three isomeric hexanes, namely n -hexane, 2-methylpentane and 3-methylpentane. The branching index (BI) for each is calculated by multiplying the number of non-hydrogen bonds made by each atom with the number on an adjacent atom, taking the reciprocal square root of the product, then summing across all non-hydrogen atoms.

Hence, for n -hexane the BI is $1/\sqrt{2} + 1/\sqrt{4} + 1/\sqrt{4} + 1/\sqrt{4} + 1/\sqrt{2}$, or 2.914. The BIs for 2-methylpentane and 3-methylpentane are 2.770 and 2.808 respectively. Clearly the Randić index can readily differentiate between alkane isomers.

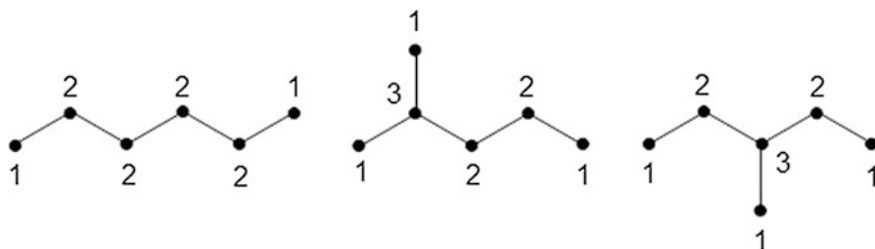


Fig. 2 Hydrogen-suppressed graphs of three isomeric hexane molecules

This was an important development, and was recognised as such by Lemont (Monty) Kier and Lowell Hall, who proceeded to develop Randić's concept and widen its applications (see Sect. 16). Randić himself later (Randić 2001) generously acknowledged the significant contributions of Kier and Hall, quoting Wilson (1952): "Every once in a while some new theory or a new experimental method or apparatus makes it possible to enter a new domain. Sometimes it is obvious to all that this opportunity has arisen, but in other cases recognition of the opportunity requires more imagination".

14 Molecular Connectivity Indices

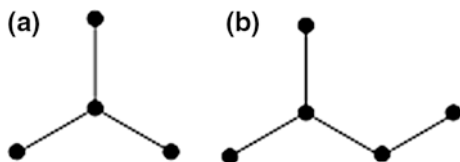
Kier and Hall, discerning the potential of the Randić index, collaborated with Randić (Kier et al. 1975) to show that his branching index could be applied to compounds other than alkanes, and could be used in QSPR and QSAR modeling. They demonstrated that the cavity surface area of 69 alcohols and hydrocarbons was well modelled by BI, which they chose to call the molecular connectivity index, χ ($r^2 = 0.956$, $s = 11.2$). Such a correlation is not unexpected, since χ clearly increases with molecular size. What is much more interesting is that they also correlated a biological activity (minimum blocking concentration of non-specific local anaesthetics) with χ for a chemically diverse set of 36 chemicals ($r^2 = 0.966$, $s = 0.390$).

It was also realised (Murray et al. 1975) that for compounds with π -bonds, better correlations could be achieved if a π -bond was regarded as two bonds, so that, for example, $\text{CH}_2 =$ has a vertex value (connectivity) of 2. Kier et al. (1976) also introduced the concept of an expanded series of the molecular connectivity index χ , involving calculation across more than one bond. For example, for isopentane, the first-order χ value, designated as $^1\chi$, is $1/\sqrt{3} + 1/\sqrt{3} + 1/\sqrt{6} + 1/\sqrt{2}$, or 2.270. The second-order χ value, calculated by multiplying across two bonds, and designated as $^2\chi$, is $1/\sqrt{1} \times 3 \times 1 + 1/\sqrt{1} \times 3 \times 2 + 1/\sqrt{1} \times 3 \times 2 + 1/\sqrt{3} \times 2 \times 1$, or 1.558. The third-order χ value, calculated by multiplying across three bonds, and designated as $^3\chi$, is $1/\sqrt{1} \times 3 \times 2 \times 1 + 1/\sqrt{1} \times 3 \times 2 \times 1$, or 0.816. Higher-order χ values are calculated similarly, and the zero-order χ value is calculated as the sum of the reciprocal square roots of the vertex values.

A further step (Kier and Hall 1976a) involved the specific treatment of heteroatoms. The vertex value δ of an atom is equivalent to the number of valence electrons minus the number of hydrogen atoms bonded to it; for example, for the N atom of NH_2 , $\delta = 3$, and for the N atom of NH , $\delta = 4$. However, this does not work for the halogens. To circumvent this problem, the δ values for halogen atoms were derived from modeling of molar refraction data, yielding δ values of: fluorine -20, chlorine 0.690, bromine 0.254, iodine 0.085.

When χ values corrected for unsaturation and heteroatoms are used, they are written as χ^v . For example, the simple (uncorrected) $^1\chi$ value for vinyl chloride,

Fig. 3 **a** 3rd order cluster;
b 4th order path-cluster



$\text{CH}_2 = \text{CHCl}$, is $1/\sqrt{2} + 1/\sqrt{2}$, or 1.414. The valence-corrected ${}^1\chi^v$ value is $1/\sqrt{2} \times 3 + 1/\sqrt{3} \times 0.690$, or 1.103.

Also in 1976 Kier and Hall (1976b) published their first book on molecular connectivity.

When branching occurs in a molecule, the atoms at and around the branch are termed a cluster (Fig. 3a) or a path-cluster (Fig. 3b). Clearly their molecular connectivity terms (${}^3\chi_c$ and ${}^4\chi_{pc}$) describe local structural properties. Kier et al. (1977) found them useful in modelling odorants (Eq. 8) and Sabljic and Protić-Sabljić (1983) used them to model properties of branched alcohols.

$$\begin{aligned} \text{Odor similarity} &= 7.47 - 1.84^2\chi + 1.34^3\chi_c \\ n &= 15 \quad r^2 = 0.848 \quad s = 0.395 \end{aligned} \quad (8)$$

Ring (termed chain) molecular connectivities describe the types of rings and their substitution patterns in a molecule (Kier and Hall 1986). Sabljic (1985) found that a chain term was required to model chromatographic retention indices of chlorinated benzenes on a polar stationary phase:

$$\begin{aligned} I^{\text{CW20M}} &= 226.8^3\chi + 1588.0^7\chi_{\text{CH}} + 649.1 \\ n &= 13 \quad r^2 = 0.996 \quad s \text{ not given} \quad F = 1347 \end{aligned} \quad (9)$$

A differential molecular connectivity index, $\Delta\chi$, was introduced in Kier and Hall (1991), defined as the difference between the simple and valence connectivity indices of the same order. The information encoded by this differential index is largely electronic. For example, Kier and Hall (1991) found that the ionization potentials (IP) of amines, alcohols and ethers were well modeled by two $\Delta\chi$ values:

$$\begin{aligned} \text{IP} &= 5.01 \Delta^0\chi + 5.17 \Delta^1\chi + 5.34 \\ n &= 24 \quad r^2 = 0.912 \quad s = 0.30 \quad F = 109 \end{aligned} \quad (10)$$

Kier and Hall have utilized molecular connectivities to model a wide range of physicochemical and biological endpoints (Hall and Kier 1999a), from aqueous solubility (Hall et al. 1975) to muscarinic receptor affinity (Kier and Hall 1978) to fish toxicity (Hall et al. 1989).

Hall and Kier (1999a) have also pointed out that molecular connectivities, together with kappa and E-state indices (see Sects. 17 and 21), have been utilized in

database characterization (Cummins et al. 1996) and combinatorial library design (Zheng et al. 1998a, b).

It is not surprising that there is considerable collinearity of χ values, especially amongst the lower order values. It is important to eliminate collinearity of descriptors in a QSAR model, otherwise distortion of the statistics can occur (Dearden et al. 2009) and mechanistic interpretation is difficult. Murcia-Soler et al. (2001) used χ values to model the anti-hyperglycemic effect and other properties of sulfonyleurea drugs. They modeled the plasma protein binding of those drugs with three molecular connectivities, namely ${}^0\chi^v$, ${}^1\chi$ and ${}^1\chi^v$, all of which are highly correlated with each other ($r > 0.99$). Lu et al. (1999) modeled the fish bioconcentration factor (BCF) of organic pollutants, and reported the following model:

$$\log \text{BCF} = 0.770 + 0.757{}^0\chi^v - 2.650{}^1\chi + 3.372{}^2\chi - 1.186{}^2\chi^v - 1.807{}^3\chi_c \quad (11)$$

$$n = 80 \quad r^2 = 0.907 \quad s = 0.364$$

They did not give the χ values of the compounds, but by comparison with the Murcia-Soler data above it is clear that at least one pair of descriptors (${}^0\chi^v$ and ${}^1\chi$) in Eq. 11 must be very highly correlated.

Randić (2001) took a different view of collinearity. He stated: “Descriptors that show high collinearity with already selected descriptors are often eliminated from structure-property-activity studies. They should not be. The *only* useful criterion for discarding a descriptor is its inability to reduce the standard error of the regression. For example, in several applications of connectivity indices, the second order connectivity index ${}^2\chi$ has been discarded because...it shows close parallelism to the connectivity index ${}^1\chi$. But... ${}^2\chi$, despite its parallelism to ${}^1\chi$, also *complements* it. That is, a part of ${}^2\chi$ which is different from ${}^1\chi$ (and which may be small) suffices to produce satisfactory regression”. Randić (2001) pointed out that a referee disagreed with his view, stating that “such a model is generally not predictive, that is, when new compounds are predicted, their presence essentially alters the interrelation between the two descriptors, ${}^1\chi$ and ${}^2\chi$ in this example. Often when models using inter-correlated variables are used, they do not produce good validation statistics”. This is confirmed by, for example, Livingstone (1995) and Hansch et al. (1998). The latter authors pointed out that a QSAR developed by Ribo and Kaiser (1984) for the toxicity of chloroanilines to *Photobacterium phosphoreum* (now called *Vibrio fischeri*) contained two highly correlated terms, ClogP and the Hammett constant σ :

$$\log 1/C = 1.25(\pm 0.49) \text{ClogP} - 1.45(\pm 1.1)\sigma + 2.01(\pm 0.70) \quad (12)$$

$$n = 14 \quad r^2 = 0.917$$

The correlation between ClogP and σ is $r^2 = 0.946$, and the σ term has low significance, as can be seen from its standard error's being high relative to its coefficient.

The present author remains of the belief that it is probably better not to use highly correlated descriptors in a QSAR/QSPR. The subject is, however, worthy of further investigation. Hollas et al. (2005) have reported that by slight modification of TIs such as molecular connectivities, the Platt number and Zagreb indices, their mutual correlation can be reduced or completely eliminated.

Another point that perhaps requires further examination is just what is meant by “highly correlated”. There does not appear to be any definitive value, at least so far as QSAR is concerned, and a wide range of values are in use. The present author’s very subjective choice is $r^2 \geq 0.8$. Gramatica et al. (2007) used $r^2 \geq 0.98$, which aligns with the view of Randić (2001). On the other hand, r values as low as 0.2 have been used as a cut-off point (Randić 2015).

It has already been mentioned that molecular connectivities, like other topological descriptors, are difficult of interpretation. Nonetheless a number of attempts have been made to do so. Kier and Hall (1976b, 1986) pointed out that there are five general categories of molecular structure described by χ indices: (i) degree of branching, (ii) variable branching pattern, (iii) position and effect of heteroatoms, (iv) adjacency patterns, and (v) degrees of cyclicality. Kier and Hall (2000, 2001) also showed that χ indices represent the numerical possibilities of a molecule encountering another identical molecule. By converting bond accessibility into a cellular automata rule for 38 alkanes, and running the dynamics, they showed that the number of cell encounters correlated better ($r^2 = 0.991$) than did $^1\chi$ ($r^2 = 0.984$) with the boiling points of the alkanes.

Randić and co-workers (Randić and Zupan 2001; Randić et al. 2001) considered the interpretation of several topological indices. They pointed out that the paucity of papers on the subject suggested that the interpretation of topological indices may be rather difficult. Nonetheless they attempted to do so, and commented that the fact that peripheral bonds make larger contributions to $^1\chi$ than do inside bonds indicates their contribution to molecular surface area, which is a measure of molecular size.

Estrada (2002) also identified χ indices as components of molecular accessibility. He interpreted the δ values (inverse square roots of the vertex degree (number of non-hydrogen bonds formed)) as the length of the arc in the van der Waals circumference accessible from outside. Then, for a $^2\chi$ index (i.e. over 3 atoms), the 3 δ values are multiplied together to give a molecular accessibility volume.

In a principal component (PC) analysis of 108 *n*-alkanes and polychlorinated biphenyls, Burkhardt et al. (1983) found that three PCs accounted for 98% of the variance in the data set, and that those PCs were associated with, respectively, (i) degree of branching, (ii) molecular size or bulkiness, and (iii) structural flexibility.

Dearden et al. (1988) looked at the correlations between a range of χ values of 59 substituents attached to a benzene ring and 54 non- χ properties of each of the substituents. In general, they found that path connectivity terms, of whatever order (≤ 6) and whether simple or valence-corrected, model predominantly bulk volume.

The $(\chi - \chi^v)$ terms did not appear to model any of the 54 non- χ properties. Other ad hoc comments have occasionally appeared in the literature; for example, Krishnasamy et al. (2008) stated that ${}^4\chi_p^v$ highlights the role of molecular surface. Stankevich et al. (1995) reported that χ indices for conjugated hydrocarbons correlated with the Hamiltonian function describing the π -electron properties of the compounds.

There have also been several attempts to “correct” χ indices. For example Li et al. (2003) developed a novel valence χ value and found it better ($r^2 = 0.939$) than the standard χ value ($r^2 = 0.889$) for the prediction of aqueous solubility of a diverse group of 36 organic compounds. Zhang et al. (2005) also used the same novel valence χ value, together with several quantum-chemical descriptors, in the modeling of corrosion inhibitory activity of 34 compounds such as imidazoles and imidazolines.

Dearden et al. (2004) devised an approach to improve the correlation of χ values with hydrophobicity by subtracting, instead of adding, the bond contributions (δ values) for bonds where one of the atoms is a heteroatom other than halogen, to give a ${}^1\chi^p$ value. For example, for a set of 23 diverse substituents, the correlation between ${}^1\chi$ and π (the hydrophobic substituent constant) was $r^2 = 0.123$, whilst that between ${}^1\chi^p$ and π was $r^2 = 0.771$.

15 Kappa Indices

Kier (1985) devised a numerical index (kappa, κ) of molecular shape from the hydrogen-suppressed graph of a molecule. It is based on the count of 2-bond fragments in a graph relative to the maximum number possible (if the molecule is star-shaped) and the minimum number in the isomeric linear graph. He showed that the sweet taste potency of 14 nitro- and cyano-anilines was modelled better ($r^2 = 0.852$, $s = 0.30$) with ${}^2\kappa$ and $({}^2\kappa)^2$ than was found by Iwamura (1980) using a Verloop steric constant ($r^2 = 0.810$, $s = 0.32$). Kier later (Kier 1986a) introduced different orders of kappa indices and (Kier 1986b) a modification term α for non-carbon atoms.

Solomon et al. (2009) used a first-order κ index in a good 5-descriptor QSAR to model the butylcholinesterase inhibition of 59 *N*-aryl derivatives ($r^2 = 0.884$).

16 Flexibility Indices

Almost all organic molecules are flexible, and flexibility often plays an important part in chemical reactions, and in xenobiotic transport and receptor binding within an organism (Luisi 1977). Kappa indices were used by Kier (1989) to develop a

molecular flexibility index. The heteroatom-weighted kappa indices ${}^1\kappa_\alpha$ and ${}^2\kappa_\alpha$ are obtained, and the flexibility index Φ is defined as:

$$\Phi = {}^1\kappa_\alpha \cdot {}^2\kappa_\alpha / A \quad (13)$$

where A = atom count.

The compressibility of a molecule is a function of the free space between molecules, which must relate to molecular flexibility. Kier and Hall (1999) reported that, for a heterogeneous set of cyclic and acyclic hydrocarbons, compressibility (K_T) correlated well with their Φ values:

$$\begin{aligned} K_T &= 17.785 \Phi + 75.032 \\ n &= 10 \quad r^2 = 0.922 \quad s = 9.4 \end{aligned} \quad (14)$$

Melting point is a function of crystal packing, which also would be expected to relate to molecular flexibility. Eike et al. (2003) obtained a good QSPR for the melting points of 75 quaternary ammonium salts with acyclic saturated alkyl side-chains, using 5 descriptors including Φ ($r^2 = 0.775$).

17 The Variable Connectivity Index

Topological indices are known as graph invariants. However, Randić (1991a, b) introduced the concept of optimization, by using weighted path numbers, of such indices involving heteroatoms in order to obtain better QSAR/QSPR models. The topic then lay dormant until Randić and Basak (1999), Krenkel et al. (2001), Pompe et al. (2004), Randić et al. (2004) and Mu et al. (2009) extended it. Singh et al. (2014) devised some variable Zagreb indices with high discriminating power. Randić and Basak (1999) were able to show that the use of two optimized variable connectivity indices improved QSPR modeling of the boiling points of 58 aliphatic alcohols, with the standard error of prediction being lowered from 6.64° to 3.89°.

Randić (2015) made the impressive point that, in the prediction of boiling points of 100 alcohols, a single variable first-order connectivity index yielded $r^2 = 0.982$, $s = 4.21^\circ$, whereas four non-variable connectivity indices were required to achieve similar statistics ($r^2 = 0.982$, $s = 4.91^\circ$) using the same data.

Randić et al. (2004) have, however, pointed out that as the training set of compounds is changed, the values of the variable indices also change. This means that the method is not fully transferable. It is also likely that unless external test set compounds are very similar to those in the training set, poor external predictivity could result. Additionally there could be a risk that the standard error of prediction could be significantly lower than the experimental error, which is unacceptable (Dearden et al. 2009).

18 Use of Topological Descriptors in Inverse QSAR

It is relatively easy, given a data set of biological activities or physicochemical properties, to describe them quantitatively with a QSAR/QSPR model. But if such a model is to be used predictively, for example to develop more potent drugs, how does one obtain potential drug candidates from the model? The problem is, of course, not restricted to models using topological indices, but Kier and Hall have examined it from that standpoint (Kier et al. 1993a, b; Hall and Kier 1993; Kier and Hall 1993).

Essentially, the approach is (Hall and Kier 1993): (i) set desired target range for property value; (ii) use QSAR equation(s) to obtain target range of each χ index; (iii) convert χ target range into target range of path count; (iv) obtain target range for number of atoms and rings; (v) use interconversion equations to obtain target degree sets; (vi) convert each degree set into a set of corresponding graphs, called candidate graphs; (vii) use best QSAR equation to predict property value for each candidate graph. Kier and Hall (1993) used the above algorithm to design potential isonarcotic agents from a published data set, and found 8 compounds likely to have isonarcotic activity in the desired range.

The Zefirov group has also examined the inverse QSAR problem with the use of a number of topological indices (Baskin et al. 1989; Gordeeva et al. 1990; Zefirov et al. 1991), using an approach quite similar to that of Kier and Hall. Skvortsova et al. (1992, 1993) similarly examined the inverse QSAR problem with the use of kappa indices.

19 Electrotological State Indices

Almost all molecular descriptors encode essentially either electronic or topological information, and usually represent whole molecules (Hall et al. 1991a). In the early 1990s Kier and Hall developed a set of descriptors that encompass both electronic and topological features, and are atom-based (Kier and Hall 1990; Hall et al. 1991a, b; Hall and Kier 1995); they termed these electrotological state (E-state) indices. This work was later drawn together in a book (Hall and Kier 1999b).

The electronic factor is considered to relate to the count of non- σ (π and lone-pair) electrons associated with an atom, and is equal to $(\delta^v - \delta)$, where δ^v is the count of valence electrons and δ is the count of σ electrons. The atom intrinsic factor I is defined as $(\delta^v + 1)/\delta$ for first row atoms, and for higher level atoms as $[(2/N)^2 \cdot \delta^v + 1]/\delta$. The perturbation ΔI_i of other atoms j on atom i is defined as:

$$\Delta I_j = \sum_{j=1}^N (I_i - I_j) / r_{ij}^2 \quad (15)$$

where N = principal quantum number, and r = count of atoms in the shortest path connecting atoms i and j , counting both i and j . Hence the E-state value S_i for atom i is $(I_i + \Delta I_i)$.

The power and beauty of the E-state approach are that, unlike most molecular descriptors, it allows an investigator to focus on the effects of individual atoms within a molecule on the activity or property under investigation, and thus potentially aids in determination of mechanism of action. Hall et al. (1991a) found that the ^{17}O NMR chemical shifts for a series of 10 ethers was modelled almost as well by an E-state term ($r^2 = 0.990$, $s = 4.3$) as by a quantum mechanically calculated partial charge ($r^2 = 0.994$, $s = 3.4$).

Another example of E-state correlation with a physicochemical property was given by Hall and Kier (1995), who modeled the boiling points (BP) of 245 alkanes and alcohols:

$$\begin{aligned} \text{BP} &= 8.21 \text{ SsCH}_3 + 14.86 \text{ SssCH}_2 + 24.56 \text{ SsssCH} + 43.76 \text{ SssssC} + 11.63 \text{ SsOH} - 43.95 \\ n &= 245 \quad r^2 = 0.941 \quad s = 8.0 \quad F = 755 \end{aligned} \tag{16}$$

Note that lower case s indicates the number of non-hydrogen bonds formed by each type of atom.

E-state indices were used by Huuskonen et al. (1999) to model the octanol-water partition coefficients ($\log P$) of 300 drugs and related compounds, yielding $r^2 = 0.87$, $s = 0.68$. However, a total of 19 E-state values were required in order to achieve those statistics. Whilst that probably reflected the diversity of the data set, it is not comparable with the results obtained by Abraham et al. (1994), who modeled $\log P$ of 613 diverse organic compounds with only four descriptors, yielding $r^2 = 0.995$, $s = 0.116$.

Kellogg et al. (1996) then introduced E-state values for hydrogen ($I(\text{H})$), mainly to take account of hydrogen bonding. They assumed $I(\text{H})$ to be dependent primarily on the attached atom, and calculated it as $I(\text{H}) = (\delta^v - \delta)^2/\delta$. Rose et al. (2002) found these E-state terms valuable for modeling blood-brain barrier partitioning of 102 diverse compounds, using two hydrogen E-state terms and a χ difference term, with reasonable statistics ($r^2 = 0.66$, $s = 0.45$).

Numerous other workers have found E-state descriptors to be of value in QSAR/QSPR modeling. Ray et al. (2010) used a combination of these and physicochemical descriptors to model the free-radical scavenging activity of 36 hydroxyphenylureas, with $q^2 = 0.957$ and external predictive $r^2 = 0.966$. An exploration of the pharmacophore of some benzodiazepine derivatives as anti-Alzheimer agents was performed by Debnath et al. (2004) using E-state descriptors, with excellent results.

Roy and Mitra (2012) have recently reviewed the use of E-state indices in drug design, property prediction and toxicity assessment. They commented that: "the... use of E-state parameters in the field of computational chemistry portray(s) them as an indispensable tool to expedite investigation of molecular mechanisms and

rational design of molecules, in addition to characterization of physicochemical properties of the molecules and identification of toxic industrial wastes and environmental pollutants”. The present author concurs with those sentiments. In a recent mechanism-based study of compounds causing skin sensitization, Dearden et al. (2015a) found, in 8 QSAR models selected by step-wise regression, that χ values, E-state indices and Kier flexibility terms featured strongly.

20 Biodescriptors

As interest in, and knowledge of, genomics and proteomics increase apace, the biological information available is huge. For example, a proteomics map derived from 2D gel electrophoresis can yield data on the charge, mass and abundance of about 2000 individual proteins (Basak and Gute 2008). The question thus arises as to whether graph theoretical indices can be devised for the characterization of biological data such as DNA sequences or proteomics maps (Basak and Gute 2008). Nandy and Basak (2000) and, Randić et al. (2000) were the first to attempt this. Nandy and Basak examined the effect of toxic substances on DNA primary sequences, and developed simple numerical descriptors from a graphical representation technique that enabled easy visualization of changes in base mutations and deletions arising from toxicity. Nandy et al. (2006) have compared a number of different approaches. The lack of correspondence amongst them led the authors to comment that “until a reasonably dependable characterization system is developed, the underlying graphical systems to be used should be the ones with intuitive appeal to understand the base composition and distribution structure in a sequence, and develop numerical techniques based on such graphs”. Basak and Gute (2008) examined several approaches to the development of mathematical biodescriptors, and concluded that they have a reasonable ability to distinguish between proteomics patterns that result from closely related chemicals and complex mixtures. This could allow the development of new drug candidates, and also act as early warning signals of toxicity. Basak (2010) has reviewed the field up to 2010.

21 Chirality

Graph theoretical indices are 2D descriptors, and so generally cannot distinguish between 3D structural features such as chirality, although Randić et al. (1990) reported an extension of the Randić index approach to give 3D descriptors. A number of attempts have been made to develop TIs that can differentiate between diastereoisomers (Golbraikh et al. 2001) and also between *cis* and *trans* enantiomers (Schultz et al. 1995). Natarajan et al. (2007) have discussed these, and developed a novel approach using a three-point interaction model, whereby the three groups of highest priority attached to a chiral center are viewed from a given

reference point. Attached groups/atoms are assigned δ values by the Kier and Hall method (Kier and Hall 1986), and decreasing importance with increasing topological distance is assigned. The group δ_i value is then calculated as follows:

$$\delta_i^v = \delta_{n1}^v + (\delta_{n2}^v/2) + (\delta_{n3}^v/4) + (\delta_{n4}^v/8) + \dots \quad (16)$$

The relative chirality index (v RCI) is then calculated as the sum of the δ_i^v values across all relevant bonds in the chiral molecule.

They found that their approach gave good differentiation between diastereoisomers and between enantiomers.

22 Software for Calculation of Topological Indices

There is now a wide range of software available for the calculation of topological indices, and many of these are listed in Table 1. It is sometimes difficult to ascertain whether or not a given software package will calculate a particular topological index, as some websites are not very specific. A number of software websites state, for example, simply that their software will calculate “topological descriptors”, without saying which ones, and such software has not been included in Table 1.

Another matter of potential concern is the accuracy of the topological indices calculated by available software packages, since a number of such programs will probably have been written in-house. A case in point is a paper by Murcia-Soler et al. (2001), who modeled anti-diabetic potencies of drugs using molecular connectivities calculated by their own in-house software. Dearden et al. (2015b) found, in a re-investigation of the Murcia-Soler data, that their reported χ values were incorrect. For example, their $^0\chi^v$ values were all too low by 0.587, and their $^3\chi^v$ values were all too low by 0.230, in comparison with the Molconn-Z values of Hall and Kier, which one would expect to be correct.

23 Conclusions

It is clear from what is written above that there is now a vast literature on the development of topological indices, and on their applications in QSAR and QSPR. One question that has arisen more than once is: are there now enough (or more than enough) topological indices available? It is true that we now have a wide range of TIs available for use in QSAR/QSPR, as this chapter shows. However, one could ask the same question regarding other descriptors, of which we now have thousands (Todeschini and Consonni 2009), yet it does not appear to have been asked, or at least not to the same extent.

Another often-voiced criticism of TIs is that they are difficult to interpret. Kubinyi (1993) implied that he regarded them as “an irrelevance which has had the

unfortunate effect of diverting attention from the real work that needs doing”. Unger (1987), reviewing Kier and Hall’s (1986) book, stated that molecular connectivity “is a highly concocted piece of numerology and is often applied with total lack of rationale”. However, Randić et al. (2016) have made the valid point that although the use of physicochemical properties in QSAR modelling can offer insights into mechanisms of action, they simply relate to parallelisms between such properties and activity, but they tell us nothing about the *structure*-activity relationship directly.

In the present author’s opinion, topological and physicochemical descriptors should be regarded as complementary. Topological indices are one, or perhaps more than one, class of QSAR/QSPR descriptors. Part of the concern about their use is that, on the whole, they have been used as stand-alone descriptors, perhaps even in competition with other types of descriptor, leading to inter-necine rivalry and hence argument: “My descriptors are better than your descriptors”. But why should this be so? All descriptors, of whatever nature and derivation, contain information that could be valuable in modeling, so why not use a descriptor pool of various types of descriptor, as proposed by Basak et al. (1999)? This point has been made strongly by Tseng et al. (2012): “There is no logical reason for keeping descriptor classes segregated. Certainly one can appreciate situations, based upon the endpoint of interest, where multiple classes of descriptors are needed to adequately capture the molecular features and interactions that contribute to the endpoint of interest”. The present author concurs fully with those sentiments. Randić (2008) noted the continuing hostility towards chemical graph theory, and has quoted verbatim many adverse comments by authors and journal editors, although he commented that molecular connectivities have been under less attack of recent years. The present author has witnessed at first hand the verbal abuse of molecular connectivity work in the presence of Kier and Hall at international conferences.

Randić (2008) commented that graph theory is widely appreciated and acknowledged in physics and biology, so why not in chemistry? Who is afraid of graph theory? He also cited many publications in which chemical graph theory has been unfairly attacked, in his view. Randić (2008) put this hostility down to ignorance of the power of graph theory, and speculated as to whether the blindness of critics could “reflect conscious or unconscious concerns how to preserve the monopoly in an applied area of medicinal and physical chemistry”. He expressed similar concerns earlier (Randić 2001) and in a recent co-authored book (Randić et al. 2016). Prelog (1976) has pointed out that “pictorial representations of graphs are so easily intelligible that chemists are often satisfied with inspecting and discussing them without paying too much attention to their algebraic aspects, but it is evident that some familiarity with the theory of graphs is necessary for deeper understanding of their properties”.

It is acknowledged that topological indices are often difficult to interpret in physicochemical terms, although it has to be said that the same applies to a great many non-topological descriptors. A number of authors have attempted such interpretation, and a few such are Basak et al. (1987, 2015), Stankevich et al. (1995), Kier and Hall (2000), Randić et al. (2001), Randić and Zupan (2001) and

Shafiei (2015). Basak (2013a) has also presented a philosophical view of mathematical chemistry.

Although most of the work reported here relates to the application of topological indices in QSAR and QSPR modeling, TIs are also being used in other fields such as proteomics and DNA sequencing (see Sect. 20) and molecular similarity (Basak et al. 1988, 2006; Randić 2014). The latter has potential for database characterization (Cummins et al. 1996) and combinatorial library design (Zheng et al. 1998a, b).

It therefore seems that the future of topological indices and their application to chemistry, biochemistry, biology and medicine are assured for the foreseeable future. There will no doubt continue to be disagreements, but that is the nature of science. If one goes back to 19th century scientific publications, it is clear that bitter arguments were taking place even then.

Let us be grateful that the dire predictions of Auguste Comte (1798–1857) did not come to pass. He wrote: “Every attempt to employ mathematical methods in the study of chemical questions must be considered profoundly irrational and contrary to the spirit of chemistry. If mathematical analysis should ever hold a prominent place in chemistry—an aberration which is happily almost impossible—it would occasion a rapid and widespread degeneration of that science” (Liang et al. 1993).

References

- Abraham, M. H., Chadha, H. S., & Mitchell, R. C. (1994). Hydrogen bonding. Part 32. An analysis of water-octanol and water-cyclohexane partitioning and the log P parameter of Seiler. *Journal of Pharmaceutical Sciences*, *83*, 1085–1100.
- Abreu, R. M. V., Ferreira, I. C. F. R., & Queiroz, M. J. R. P. (2009). QSAR model for predicting radical scavenging activity of di(hetero)arylamines derivatives of benzo[b]thiophenes. *European Journal of Medicinal Chemistry*, *44*, 1952–1958.
- Aptula, A. O., Jeliaskova, N. G., Schultz, T. W., & Cronin, M. T. D. (2005). The better predictive model: high q^2 for the training set or low root mean square error of prediction for the test set? *QSAR & Combinatorial Science*, *24*, 385–396.
- Bajaj, S., Sami, S. S., & Madan, A. K. (2005). Prediction of anti-inflammatory activity of *N*-arylanthranilic acids: Computational approach using refined Zagreb indices. *Croatica Chemica Acta*, *78*, 165–174.
- Balaban, A. T. (1979). Chemical graphs. 34. Five new topological indices for the branching of tree-like graphs. *Theoretica Chimica Acta*, *53*, 355–375.
- Balaban, A. T. (1982). Highly discriminating distance-based topological index. *Chemical Physics Letters*, *89*, 399–404.
- Balaban, A. T. (1985). Applications of graph theory in chemistry. *Journal of Chemical Information and Computer Sciences*, *25*, 334–343.
- Basak, S. C. (1987). Use of molecular complexity indices in predictive pharmacology and toxicology: A QSAR approach. *Medical Science Research*, *15*, 605–609.
- Basak, S. C., Magnuson, V. R., Niemi, G. J., Regal, R. R., & Veith, G. D. (1987). Topological indices: Their nature, mutual relatedness, and applications. *Mathematical Modelling*, *8*, 300–305.
- Basak, S. C., Magnuson, V. R., Niemi, G. J., & Regal, R. R. (1988). Determining structural similarity of chemical using graph-theoretic indices. *Discrete Applied Mathematics*, *19*, 17–44.

- Basak, S. C. (1999). Information theoretic indices of neighborhood complexity and their applications. In J. Devillers & A. T. Balaban (Eds.), *Topological indices and related descriptors in QSAR and QSPR* (pp. 563–593). Amsterdam: Gordon and Breach Science Publishers.
- Basak, S. C., Gute, B. D., & Grunwald, G. D. (1999). A hierarchical approach to the development of QSAR models using topological, geometrical and quantum chemical parameters. In J. Devillers & A. T. Balaban (Eds.), *Topological indices and related descriptors in QSAR and QSPR* (pp. 563–593). Amsterdam: Gordon and Breach Science Publishers.
- Basak, S. C., Balaban, A. T., Grunwald, G. D., & Gute, B. D. (2000). Topological indices: Their nature and mutual relatedness. *Journal of Chemical Information and Computer Sciences*, *40*, 891–898.
- Basak, S. C., Gute, B. D., & Mills, D. (2006). Similarity methods in analog selection, property estimation and clustering of diverse chemicals. *ARKIVOC*, *1*, 157–210.
- Basak, S. C., & Gute, B. D. (2008). Mathematical biodescriptors of proteomics maps: Background and applications. *Current Opinion in Drug Discovery & Development*, *11*, 320–326.
- Basak, S. C. (2010). Role of mathematical chemodescriptors and proteomics-based biodescriptors in drug discovery. *Drug Development Research*, *72*, 1–9.
- Basak, S. C. (2013a). Philosophy of mathematical chemistry: A personal perspective. *HYLE—International Journal for Philosophy of Chemistry*, *19*, 4–17.
- Basak, S. C. (2013b). Mathematical descriptors for the prediction of property, bioactivity, and toxicity of chemicals from their structure: A chemical-cum-biochemical approach. *Current Computer-Aided Drug Design*, *9*, 449–462.
- Basak, S. C., Grunwald, G. D., & Majumdar, S. (2015). Intrinsic dimensionality of chemical space: Characterization and applications. *Mol2Net Section B, Proceedings*, *1*, 1–10.
- Baskin, I. I., Gordeeva, E. V., Devdarian, R. O., Zefirov, N. S., Palyulin, V. A., & Stankevich, M. I. (1989). Methodology of solution of the inverse problem for the structure-property relationship for the case of topological indices. *Doklady Chemistry*, *307*, 217–220.
- Bharate, S. B., & Singh, I. P. (2011). Quantitative structure-activity relationship study of phloroglucinol-terpene adducts as anti-leishmanial agents. *Bioorganic & Medicinal Chemistry Letters*, *21*, 4310–4315.
- Burkhard, L. P., Andren, A. W., & Armstrong, D. E. (1983). Structure activity relationships using molecular connectivity indices with principal component analysis. *Chemosphere*, *12*, 935–943.
- Caballero, J., Fernández, M., Saavedra, M., & González-Nilo, F. D. (2008). 2D autocorrelation, CoMFA, and CoMSIA modeling of protein tyrosine kinases' inhibition by substituted pyrido [2,3-*d*]pyrimidine derivatives. *Bioorganic & Medicinal Chemistry*, *16*, 810–821.
- Cronin, M. T. D., et al. (2002). Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to *Tetrahymena pyriformis*. *Chemosphere*, *49*, 1201–1221.
- Cummins, D. J., Andrews, C. W., Bentley, J. A., & Cory, M. (1996). Molecular diversity and chemical databases: Comparison of medicinal chemistry knowledge bases and databases of commercially available compounds. *Journal of Chemical Information and Computer Sciences*, *36*, 750–763.
- Dearden, J. C., Bradburne, S. J. A., Cronin, M. T. D., Solanki, P., et al. (1988). The physical significance of molecular connectivity. In J. E. Turner, M. W. England, T. W. Schultz, & N. J. Kwaak (Eds.), *QSAR-88* (pp. 43–50). Oakridge, TN: U.S. Department of Energy.
- Dearden, J. C., Cronin, M. T. D., Schultz, T. W., & Lin, D. T. (1995). QSAR study of the toxicity of nitrobenzenes to *Tetrahymena pyriformis*. *Quantitative Structure-Activity Relationships*, *14*, 189–196.
- Dearden, J. C., Wong, E. H. Y., & Walker, J. D. (2004). Molecular connectivity: polarity correction to improve correlation with hydrophobicity. *QSAR & Combinatorial Science*, *23*, 75–79.
- Dearden, J. C., Cronin, M. T. D., & Kaiser, K. L. E. (2009). How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR and QSAR in Environmental Research*, *20*, 241266.
- Dearden, J. C., et al. (2015a). Mechanism-based QSAR modeling of skin sensitization. *Chemical Research in Toxicology*, *28*, 1975–1968.

- Dearden, J. C., Hewitt, M., & Rowe, P. H. (2015b). QSAR study of some anti-hyperglycaemic sulphonylurea drugs. *SAR & QSAR in Environmental Research*, *26*, 439–448.
- Debnath, B., Gayen, S., Basu, A., Srikanth, K., & Jha, T. (2004). Quantitative structure-activity relationship study on some benzodiazepine derivatives as anti-Alzheimer agents. *Journal of Molecular Modelling*, *10*, 328–334.
- Devillers, J. (1999a). No-free-lunch molecular descriptors in QSAR and QSPR. In J. Devillers & A. T. Balaban (Eds.), *Topological indices and related descriptors in QSAR and QSPR* (pp. 1–20). Amsterdam: Gordon and Breach Science Publishers.
- Devillers, J. (1999b). Autocorrelation descriptors for modeling (eco)toxicological endpoints. In J. Devillers & A. T. Balaban (Eds.), *Topological indices and related descriptors in QSAR and QSPR* (pp. 595–612). Amsterdam: Gordon and Breach Science Publishers.
- Dureja, H., Gupta, S., & Madan, A. K. (2008). Topological models for prediction of pharmacokinetic parameters of cephalosporins using random forest, decision tree and moving average analysis. *ScientiaPharmaceutica*, *76*, 377–394.
- Eike, D. M., Brennecke, J. F., & Maginn, E. J. (2003). Predicting melting points of quaternary ammonium ionic liquids. *Green Chemistry*, *5*, 323–328.
- Estrada, E. (2002). Physicochemical interpretation of molecular connectivity indices. *Journal of Physical Chemistry A*, *106*, 9085–9091.
- Filip, P. A., Balaban, T. S., & Balaban, A. T. (1987). A new approach for devising graph invariants: Derived topological indices with low degeneracy and good correlation ability. *Journal of Mathematical Chemistry*, *1*, 61–83.
- Golbraikh, A., & Tropsha, A. (2002). Beware of q^2 ! *Journal of Molecular Graphics and Modelling*, *20*, 269–276.
- Golbraikh, A., Bonchev, D., & Tropsha, A. (2001). Novel chirality descriptors derived from molecular topology. *Journal of Chemical Information and Computer Sciences*, *41*, 147–158.
- González, M. P., Caballero, J., Helguera, A. M., et al. (2006). 2D autocorrelation modelling of the inhibitory activity of cytokinin-derived cyclin-dependent kinase inhibitors. *Bulletin of Mathematical Biology*, *68*, 735–751.
- Gordeeva, E. V., Molchanova, M. S., & Zefirov, N. S. (1990). General methodology and computer program for the exhaustive restoring of chemical structures by molecular connectivity indexes. *Tetrahedron Computer Methodology*, *3*, 389–415.
- Gordon, M., & Scantlebury, G. R. (1964). Non-random polycondensation: statistical theory of the substitution effect. *Transactions of the Faraday Society*, *60*, 604–621.
- Gramatica, P., Pilutti, P., & Papa, E. (2007). Approaches for externally validated QSAR modelling of nitrated polycyclic aromatic hydrocarbon mutagenicity. *SAR and QSAR in Environmental Research*, *18*, 169–178.
- Gupta, M. K., Sagar, R., Shaw, A. K., & Prabhakar, Y. S. (2005). CP-MLR directed QSAR studies on the antimycobacterial activity of functionalized alkenols—Topological descriptors in modeling the activity. *Bioorganic & Medicinal Chemistry*, *13*, 35–343.
- Gute, B. D., & Basak, S. C. (1997). Predicting acute toxicity (LC₅₀) of benzene derivatives using theoretical molecular descriptors: A hierarchical QSAR approach. *SAR & QSAR in Environmental Research*, *7*, 117–131.
- Gutman, I., & Trinajstić, N. (1972). Graph theory and molecular orbitals. Total π -electron energy of alternant hydrocarbons. *Chemical Physics Letters*, *17*, 535–538.
- Gutman, I., & Körtvélyesi, T. (1995). Wiener indices and molecular surfaces. *Zeitschrift für Naturforschung*, *50A*, 669–671.
- Hall, L. H., Kier, L. B., & Murray, W. J. (1975). Molecular connectivity II: Relationship to water solubility and boiling point. *Journal of Pharmaceutical Sciences*, *64*, 1974–1977.
- Hall, L. H., Maynard, E. L., & Kier, L. B. (1989). QSAR investigation of benzene toxicity to fathead minnow using molecular connectivity. *Environmental Toxicology and Chemistry*, *8*, 783–788.
- Hall, L. H., Mohney, B., & Kier, L. B. (1991a). The electrotopological state: Structure information at the atomic level for molecular graphs. *Journal of Chemical Information and Computer Sciences*, *31*, 76–82.

- Hall, L. H., Mohney, B., & Kier, L. B. (1991b). The electrotopological state: An atom index for QSAR. *Quantitative Structure-Activity Relationships*, 10, 43–51.
- Hall, L. H., & Kier, L. B. (1993). Design of molecules from quantitative structure-activity relationship models. III. Role of higher order path counts: Path three. *Journal of Chemical Information and Computer Sciences*, 33, 598–603.
- Hall, L. H., & Kier, L. B. (1995). Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. *Journal of Chemical Information and Computer Sciences*, 35, 1039–1045.
- Hall, L. H., & Kier, L. B. (1999a). Molecular connectivity chi indices for database analysis and structure-property modeling. In J. Devillers & A. T. Balaban (Eds.), *Topological indices and related descriptors in QSAR and QSPR* (pp. 307–360). Amsterdam: Gordon and Breach Science Publishers.
- Hall, L. H., & Kier, L. B. (1999b). *Molecular structure description: The electrotopological state*. San Diego: Academic Press.
- Hansch, C., Gao, H., & Hoekman, D. (1998). A generalized approach to comparative QSAR. In J. Devillers (Ed.), *Comparative QSAR* (pp. 285–368). London: Taylor & Francis.
- Hollas, B., Gutman, I., & Trinajstić, N. (2005). On reducing correlations between topological indices. *Croatica Chemica Acta*, 78, 489–492.
- Hosoya, H. (1971). Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bulletin of the Chemical Society of Japan*, 44, 2332–2339.
- Huuskonen, J. J., Villa, A. E. P., & Tetko, I. V. (1999). Prediction of partition coefficient based on atom-type electrotopological state indices. *Journal of Pharmaceutical Sciences*, 88, 229–233.
- Ivanciuc, O. (2000). QSAR comparative study of Wiener descriptors for weighted molecular graphs. *Journal of Chemical Information and Computer Sciences*, 40, 1412–1422.
- Ivanciuc, O., & Balaban, A. T. (1999). The graph description of chemical structures. In J. Devillers & A. T. Balaban (Eds.), *Topological indices and related descriptors in QSAR and QSPR* (pp. 59–167). Amsterdam: Gordon and Breach Science Publishers.
- Iwamura, H. (1980). Structure-taste relationship of perillartine and nitro- and cyanoaniline derivatives. *Journal of Medicinal Chemistry*, 23, 308–312.
- Jalali-Heravi, M., & Asadollahi-Baboli, M. (2008). QSAR analysis of platelet-derived growth inhibitors using GA-ANN and shuffling crossvalidation. *QSAR & Combinatorial Science*, 27, 750–757.
- Johnson, S. R. (2008). The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *Journal of Chemical Information and Modeling*, 48, 25–26.
- Kellogg, G. E., Kier, L. B., Gaillard, P., & Hall, L. H. (1996). The E-state fields: Applications to 3D QSAR. *Journal of Computer-Aided Molecular Design*, 10, 513–520.
- Kier, L. B., & Hall, L. H. (1976a). Molecular connectivity VII: Specific treatment of heteroatoms. *Journal of Pharmaceutical Sciences*, 65, 1806–1809.
- Kier, L. B., & Hall, L. H. (1976b). *Molecular connectivity in chemistry and drug research*. New York: Academic Press.
- Kier, L. B., Di Paolo, T., & Hall, L. H. (1977). Structure-activity studies on odor molecules using molecular connectivity. *Journal of Theoretical Biology*, 67, 585–595.
- Kier, L. B., & Hall, L. H. (1978). Molecular connectivity study of muscarinic receptor affinity of acetylcholine antagonists. *Journal of Pharmaceutical Sciences*, 67, 1408–1412.
- Kier, L. B. (1985). A shape index from molecular graphs. *Quantitative Structure-Activity Relationships*, 4, 109–116.
- Kier, L. B., & Hall, L. H. (1986). *Molecular connectivity in structure-activity analysis*. Chichester: Research Studies Press.
- Kier, L. B. (1986a). Shape indexes of orders one and three from molecular graphs. *Quantitative Structure-Activity Relationships*, 5, 1–7.
- Kier, L. B. (1986b). Distinguishing atom differences in a molecular graph shape index. *Quantitative Structure-Activity Relationships*, 5, 7–12.

- Kier, L. B. (1989). An index of molecular flexibility from kappa shape attributes. *Quantitative Structure-Activity Relationships*, 8, 221–224.
- Kier, L. B., & Hall, L. H. (1990). An electrotopological-state index for atoms in molecules. *Pharmaceutical Research*, 7, 801–807.
- Kier, L. B., & Hall, L. H. (1991). A differential molecular connectivity index. *Quantitative Structure-Activity Relationships*, 10, 134–140.
- Kier, L. B., & Hall, L. H. (1993). The generation of molecular structures from a graph-based QSAR equation. *Quantitative Structure-Activity Relationships*, 12, 383–388.
- Kier, L. B., Hall, L. H., & Frazer, J. W. (1993a). Design of molecules from quantitative structure-activity relationship models. I. Information transfer between path and vertex degree counts. *Journal of Chemical Information and Computer Sciences*, 33, 143–147.
- Kier, L. B., Hall, L. H., & Frazer, J. W. (1993b). Design of molecules from quantitative structure-activity relationship models. II. Derivation and proof of information transfer relating equations. *Journal of Chemical Information and Computer Sciences*, 33, 148–152.
- Kier, L. B., & Hall, L. H. (1999). The kappa indices for modeling molecular shape and flexibility. In J. Devillers & A. T. Balaban (Eds.), *Topological indices and related descriptors in QSAR and QSPR* (pp. 455–489). Amsterdam: Gordon and Breach Science Publishers.
- Kier, L. B., & Hall, L. H. (2000). Intermolecular accessibility: The meaning of molecular connectivity. *Journal of Chemical Information and Computer Sciences*, 40, 792–795.
- Kier, L. B., & Hall, L. H. (2001). Molecular connectivity: Intermolecular accessibility and encounter simulation. *Journal of Molecular Graphics and Modelling*, 20, 76–83.
- Kier, L. B., Hall, L. H., Murray, W. J., & Randić, M. (1975). Molecular connectivity I: Relationship to local anesthesia. *Journal of Pharmaceutical Sciences*, 64, 1971–1974.
- Kier, L. B., Murray, W. J., Randić, M., & Hall, L. H. (1976). Molecular connectivity V: Connectivity series concept applied to density. *Journal of Pharmaceutical Sciences*, 65, 1226–1230.
- Krenkel, G., Castro, E. A., & Toropov, A. A. (2001). The variable molecular descriptors based on the optimization of correlation weights of local graph invariants. *Journal of Molecular Structure (THEOCHEM)*, 542, 107–113.
- Krishnasamy, C., Raghuraman, A., Kier, L. B., & Desai, U. R. (2008). Application of molecular connectivity and electro-topological indices in quantitative structure-activity analysis of pyrazole derivatives as inhibitors of factor Xa and thrombin. *Chemistry & Biodiversity*, 5, 2609–2620.
- Kubinyi, H. (1993). *QSAR: Hansch analysis and related approaches*. Weinheim: Wiley-VCH.
- Kubinyi, H. (1997). A general view on similarity and QSAR studies. In H. van de Waterbeemd, B. Testa, & G. Folkers (Eds.), *Computer-assisted lead finding and optimization: Current tools for medicinal chemistry* (pp. 9–28). Basel, Verlag Helvetica Chimica Acta, & Weinheim, VCH.
- Li, X. H., Jalbout, A. F., & Solimannejad, M. (2003). Definition and application of a novel valence molecular connectivity index. *Journal of Molecular Structure (THEOCHEM)*, 663, 81–85.
- Liang, Y.-Z., Kvalhcim, O. M., & Manne, R. (1993). White, grey and black multicomponent systems: A classification of mixture problems and methods for their quantitative analysis. *Chemometrics and Intelligent Laboratory Systems*, 18, 235–250.
- Livingstone, D. (1995). *Data analysis for chemists: Applications to QSAR and chemical product design*. Oxford: Oxford University Press.
- Livingstone, D. J. (2000). The characterization of chemical structures using molecular properties. A survey. *Journal of Chemical Information and Computer Sciences*, 40, 195–209.
- Lopez de Compadre, R. L., Compadre, C. M., Castillo, R., & Dunn, W. J. (1983). On the use of connectivity indices in quantitative structure activity studies. *European Journal of Medicinal Chemistry*, 18, 569–571.
- Lu, X., Tao, S., Cao, J., & Dawson, R. W. (1999). Prediction of fish bioconcentration factors of nonpolar organic pollutants based on molecular connectivity indices. *Chemosphere*, 39, 987–999.
- Luisi, P. L. (1977). Molecular conformational rigidity: An approach to quantification. *Naturwissenschaften*, 64, 569–574.

- Maran, U., Sild, S., Tulp, I., Takkis, K., & Moosus, M. (2010). Molecular descriptors from two-dimensional chemical structure. In M. T. D. Cronin & J. C. Madden (Eds.), *In silico toxicology: Principles and applications* (pp. 148–192). Cambridge: RSC Publishing.
- Mekenyan, O., Bonchev, D., Sabljic, A., & Trinajstić, N. (1987). Application of topological indices to QSAR. The use of the Balaban index and the electrotopology index for correlations with toxicity of ethers on mice. *Acta Pharmaceutica Jugoslavica*, *37*, 75–86.
- Moreau, G., & Broto, P. (1980). The autocorrelation of a topological structure: A new molecular descriptor. *Nouveau Journal de Chimie*, *4*, 359–360.
- Mowshowitz, A. (1968). Entropy and the complexity of graphs: I. An index of the relative complexity of a graph. *Bulletin of Mathematical Biophysics*, *30*, 175–204.
- Mu, L., He, H., Yang, W., & Feng, C. (2009). Variable molecular connectivity indices for predicting the diamagnetic susceptibilities of organic compounds. *Industrial and Engineering Chemistry Research*, *48*, 4165–4175.
- Murcia-Soler, M., Pérez-Giménez, F., Nalda-Molina, R., Salabert-Salvador, M. T., Garcia-March, F. J., Cercós-del-Pozo, R. A., et al. (2001). QSAR analysis of hypoglycaemic agents using the topological indices. *Journal of Chemical Information and Computer Sciences*, *41*, 1345–1354.
- Murray, W. J., Hall, L. H., & Kier, L. B. (1975). Molecular connectivity III. Relationship to partition coefficients. *Journal of Pharmaceutical Sciences*, *64*, 1978–1981.
- Nandy, A., & Basak, S. C. (2000). Simple numerical descriptor for quantifying effect of toxic substances on DNA sequences. *Journal of Chemical Information and Computer Sciences*, *40*, 915–919.
- Nandy, A., Harle, M., & Basak, S. C. (2006). Mathematical descriptors of DNA sequences: Development and applications. *ARKIVOC*, *ix*, 211–238.
- Natarajan, R., Basak, S. C., & Neumann, T. S. (2007). Novel approach for the numerical characterization of molecular chirality. *Journal of Chemical Information and Modeling*, *47*, 771–775.
- Netzeva, T. I. (2004). Whole molecule and atom-based topological descriptors. In M. T. D. Cronin & D. J. Livingstone (Eds.), *Predicting chemical toxicity and fate* (pp. 61–83). Boca Raton: FL, CRC Press.
- Nikolić, S., Kovačević, G., Miličević, A., & Trinajstić, N. (2003). The Zagreb indices 30 years after. *Croatica Chemica Acta*, *76*, 113–124.
- Noorizadeh, H., Farmany, A., & Noorizadeh, M. (2011). Quantitative structure-retention relationships analysis of retention index of essential oils. *Quimica Nova*, *34*, 242–249.
- OECD (2004). Validation of (Q)SAR models. Retrieved May 26, 2016 from <http://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm>.
- Pal, D. K., Sengupta, C., & De, A. U. (1988). A new topochemical descriptor (TAU) in molecular connectivity concept: Part I—Aliphatic compounds. *Indian Journal of Chemistry*, *27B*, 734–739.
- Platt, J. R. (1947). Influence of neighbor bonds on additive bond properties in paraffins. *Journal of Chemical Physics*, *15*, 419–420.
- Pompe, M., Veber, M., Randić, M., & Balaban, A. T. (2004). Using variable and fixed topological indices for the prediction of reaction rate constants of volatile unsaturated hydrocarbons with OH radicals. *Molecules*, *9*, 1160–1176.
- Prelog, V. (1976). Foreword. In A. T. Balaban (Ed.), *Chemical applications of graph theory*. London: Academic Press.
- Randić, M. (1975). On characterization of molecular branching. *Journal of the American Chemical Society*, *97*, 6609–6615.
- Randić, M., Jerman-Blažić, B., & Trinajstić, N. (1990). Development of 3-dimensional molecular descriptors. *Computers & Chemistry*, *14*, 237–246.
- Randić, M. (1991a). Novel graph theoretical approach to heteroatoms in quantitative structure-activity relationships. *Chemometrics and Intelligent Laboratory Systems*, *10*, 213–227.
- Randić, M. (1991b). On computation of optimal parameters for multivariate analysis of structure-property relationship. *Journal of Computational Chemistry*, *12*, 970–980.

- Randić, M., & Basak, S. C. (1999). Optimal molecular descriptors based on weighted path numbers. *Journal of Chemical Information and Computer Sciences*, 39, 261–266.
- Randić, M., Vracko, M., Nandy, A., & Basak, S. C. (2000). On 3D graphical representation of DNA primary sequences and their numerical characterization. *Journal of Chemical Information and Computer Sciences*, 40, 1235–1244.
- Randić, M. (2001). The connectivity index 25 years after. *Journal of Molecular Graphics and Modelling*, 20, 19–35.
- Randić, M., & Zupan, J. (2001). On interpretation of well-known topological indices. *Journal of Chemical Information and Computer Sciences*, 41, 550–560.
- Randić, M., Balaban, A. T., & Basak, S. C. (2001). On structural interpretation of several distance related topological indices. *Journal of Chemical Information and Computer Sciences*, 41, 593–601.
- Randić, M., Pompe, M., Mills, D., & Basak, S. C. (2004). Variable connectivity index as a tool for modeling structure-property relationships. *Molecules*, 9, 1177–1193.
- Randić, M. (2008). On history of the Randić index and emerging hostility toward chemical graph theory. *MATCH Communications in Mathematical and in Computer Chemistry*, 59, 5–124.
- Randić, M. (2014). On of molecular similarity based on a single descriptor. *Chemical Physics Letters*, 599, 1–6.
- Randić, M. (2015). On the history of the connectivity index: From the connectivity index to the exact solution of the protein alignment problem. *SAR & QSAR in Environmental Research*, 26, 523–555.
- Randić, M., Novič, M., & Plavšić, D. (2016). *Solved and unsolved problems of structural chemistry*. Boca Raton, FL: CRC Press.
- Ray, S., Roy, P. P., Sengupta, C., & Roy, K. (2010). Exploring QSAR of hydroxyphenylureas as antioxidants using physicochemical and electrotopological state atom parameters. *Molecular Simulation*, 36, 484–492.
- Ribo, J. M., & Kaiser, K. L. E. (1984). Toxicities of chloroanilines to *Photobacterium phosphoreum* and their correlations and effects on other organisms and structural parameters. In K. L. E. Kaiser (Ed.), *QSAR in environmental toxicology* (pp. 319–336). Dordrecht: D. Reidel Publishing Co.
- Rose, K., Hall, L. H., & Kier, L. B. (2002). Modeling blood-brain barrier partitioning using the electrotopological state. *Journal of Chemical Information and Computer Sciences*, 42, 651–666.
- Rouvray, D. H., & Crafford, B. C. (1976). The dependence of physicochemical properties on topological factors. *South African Journal of Science*, 72, 47–51.
- Rouvray, D. H., & King, R. B. (Eds.) (2002). *Topology in chemistry*. Chichester: Horwood Publishing Limited.
- Roy, K., & Ghosh, G. (2003). Introduction of extended topochemical atom (ETA) indices in the valence electron mobile (VEM) environment as tools for QSAR/QSPR studies. *Internet Electronic Journal of Molecular Design*, 2, 599–620.
- Roy, K., & Ghosh, G. (2004). QSTR with extended topochemical atom indices. 2. Fish toxicity of substituted benzenes. *Journal of Chemical Information & Computer Sciences*, 44, 559–567.
- Roy, K., & Ghosh, G. (2009). QSTR with extended topochemical atom (ETA) indices. 12. QSAR for the toxicity of diverse aromatic compounds to *Tetrahymena pyriformis*. *Chemosphere*, 77, 999–1009.
- Roy, K., & Mitra, I. (2012). Electrotopological state atom (E-state) index in drug design, QSAR, property prediction and toxicity assessment. *Current Computer-Aided Drug Design*, 8, 135–158.
- Roy, K., Kar, S., & Das, R. N. (2015). *Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment*. Amsterdam: Academic Press.
- Roy, K., & Saha, A. (2003). QSPR with TAU indices: Water solubility of diverse functional acyclic compounds. *Internet Electronic Journal of Molecular Design*, 2, 475–491. http://biochempress.com/Files/iejmd_2003_2_0475.pdf.

- Sabljić, A. (1985). Calculation of retention indices by molecular topology: Chlorinated benzenes. *Journal of Chromatography*, *319*, 1–8.
- Sabljić, A. (1990). Topological indices and environmental chemistry. In W. Karcher & J. Devillers (Eds.), *Practical applications of quantitative structure-activity relationships (QSAR) in environmental chemistry and toxicology* (pp. 61–82). Dordrecht: Kluwer Academic Publishers.
- Sabljić, A., & Protić-Sabljić, M. (1983). Quantitative structure-activity study of the mechanism of inhibition of microsomal p-hydroxylation of aniline by alcohols. Role of steric factors. *Molecular Pharmacology*, *23*, 213–218.
- Schultz, H. P., Schultz, E. B., & Schultz, T. P. (1995). Topological organic chemistry. 9. Graph theory and molecular topological indices of stereoisomeric organic compounds. *Journal of Chemical Information & Computer Sciences*, *35*, 864–870.
- Shafiei, F. (2015). Relationship between topological indices and thermodynamic properties and of the monocarboxylic acids applications in QSPR. *Iranian Journal of Mathematical Chemistry*, *6*, 15–28.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.
- Singh, M., Das, K. C., Gupta, S., & Madan, A. K. (2014). Refined variable Zagreb indices: Highly discriminating topological descriptors for QSAR/QSPR. *International Journal of Chemical Modeling*, *6*, 403–428.
- Skvortsova, M. I., Baskin, I. I., Slovokhotova, O. L., Palyulin, V. A., & Zefirov, N. S. (1992). Inverse problem in QSAR/QSPR analysis for case of topological indexes characterizing molecular form (Kier indexes). *Doklady Chemistry*, *324*, 103–107.
- Skvortsova, M. I., Baskin, I. I., Slovokhotova, O. L., Palyulin, V. A., & Zefirov, N. S. (1993). Inverse problem in QSAR/QSPR studies for the case of topological indices characterizing molecular shape (Kier indices). *Journal of Chemical Information & Computer Sciences*, *33*, 630–634.
- Solomon, K. A., Sundararajan, S., & Abirami, V. (2009). QSAR studies on *N*-aryl derivative activity towards Alzheimer's disease. *Molecules*, *14*, 1448–1455.
- Stankevich, I. V., Skvortsova, M. I., & Zefirov, N. S. (1995). On a quantum chemical interpretation of molecular connectivity indices for conjugated hydrocarbons. *Journal of Molecular Structure (THEOCHEM)*, *342*, 173–179.
- Stiel, L. I., & Thodos, G. (1962). The normal boiling points and critical constants of saturated aliphatic hydrocarbons. *American Institute of Chemical Engineers Journal*, *8*, 527–529.
- Thakur, A., Thakur, M., Khadikar, P. V., Supuran, C. T., & Sudele, P. (2004). QSAR study on benzenesulphonamide carbonic anhydrase inhibitors: Topological approach using Balaban index. *Bioorganic & Medicinal Chemistry*, *12*, 789–793.
- Todeschini, R., & Gramatica, P. (1997). The WHIM theory: New 3D molecular descriptors for QSAR in environmental modelling. *SAR and QSAR in Environmental Research*, *7*, 89–115.
- Todeschini, R., & Consonni, V. (2009). *Molecular descriptors for chemoinformatics. Alphabetical listing* (2nd ed., Vol. 1). Weinheim: Wiley-VCH.
- Todeschini, R., Lasagni, M., & Marengo, E. (1994). New molecular descriptors for 2D and 3D structures theory. *Journal of Chemometrics*, *8*, 263–272.
- Tong, J., Liu, S., Zhou, P., Wu, B., & Li, Z. (2008). A novel descriptor of amino acids and its application in peptide QSAR. *Journal of Theoretical Biology*, *253*, 90–97.
- Tseng, Y. J., Hopfinger, A. J., & Esposito, E. X. (2012). The great descriptor melting pot: Mixing descriptors for the common good of QSAR models. *Journal of Computer-Aided Molecular Design*, *26*, 39–43.
- Unger, S. H. (1987). Molecular connectivity in structure-activity analysis (book review). *Journal of Pharmaceutical Sciences*, *76*, 269–270.
- Vlaia, V., Olariu, T., Vlaia, L., Butur, M., Ciubotariu, C., Medeleanu, M., et al. (2009). Quantitative structure-activity relationship (QSAR). IV. Analysis of the toxicity of aliphatic esters by means of weighted holistic invariant molecular (WHIM) descriptors. *Farmacia*, *57*, 511–522.
- Wells, P. R. (1968). *Linear free energy relationships*. London: Academic Press.

- Wiener, H. (1947). Structural determination of paraffin boiling points. *Journal of the American Chemical Society*, 69, 17–20.
- Wiener, H. (1948). Relation of the physical properties of the isomeric alkanes to molecular structure. *Journal of Physical Chemistry*, 52, 1082–1089.
- Wilson, E. B. (1952). *Introduction to scientific research*. New York: McGraw-Hill.
- Yilmaz, B., & Göktürk, M. (2009). Interactive data mining for molecular graphs. *Journal of Automated Methods & Management in Chemistry*. doi:[10.1155/2009/502527](https://doi.org/10.1155/2009/502527).
- Zefirov, N. S., Palyulin, V. A., & Radchenko, E. V. (1991). Problem of generation of structures with definite properties. Solution of inverse problem for Balaban centric index. *Doklady Chemistry*, 316, 921–924.
- Zhang, S. G., Lei, W., Xia, M. Z., & Wang, F. W. (2005). QSAR study of N-containing corrosion inhibitors: Quantum chemical approach assisted by topological index. *Journal of Molecular Structure (THEOCHEM)*, 732, 173–182.
- Zheng, W., Cho, S. J., & Tropsha, A. (1998a). Rational combinatorial design. 1. Focus 2-D: A new approach to the design of targeted combinatorial chemical libraries. *Journal of Chemical Information and Computer Sciences*, 38, 251–258.
- Zheng, W., Cho, S. J., & Tropsha, A. (1998b). Rational combinatorial design. 2. Rational design of targeted combinatorial peptide libraries using chemical similarity probe and inverse QSAR approaches. *Journal of Chemical Information and Computer Sciences*, 38, 259–268.

Which Performance Parameters Are Best Suited to Assess the Predictive Ability of Models?

Károly Héberger, Anita Rácz and Dávid Bajusz

Abstract We have revisited the vivid discussion in the QSAR-related literature concerning the use of external versus cross-validation, and have presented a thorough statistical comparison of model performance parameters with the recently published SRD (sum of (absolute) ranking differences) method and analysis of variance (ANOVA). Two case studies were investigated, one of which has exclusively used external performance merits. The SRD methodology coupled with ANOVA shows unambiguously for both case studies that the performance merits are significantly different, independently from data preprocessing. While external merits are generally less consistent (farther from the reference) than training and cross-validation based merits, a clear ordering and a grouping pattern of them could be acquired. The results presented here corroborate our earlier, recently published findings (SAR QSAR Environ. Res., 2015, 26, 683–700) that external validation is not necessarily a wise choice, and is frequently comparable to a random evaluation of the models.

Keywords Performance parameters (merits) · Ranking · Cross-validation · External validation · QSAR modeling

K. Héberger (✉) · A. Rácz
Plasma Chemistry Research Group, Research Centre for Natural Sciences,
Hungarian Academy of Sciences, Magyar Tudósok krt. 2, Budapest 1117, Hungary
e-mail: heberger.karoly@ttk.mta.hu

A. Rácz
e-mail: racz.anita@ttk.mta.hu

D. Bajusz
Medicinal Chemistry Research Group, Research Centre for Natural Sciences,
Hungarian Academy of Sciences, Magyar Tudósok krt. 2, Budapest 1117, Hungary
e-mail: bajusz.david@ttk.mta.hu

1 Introduction

There is a long lasting and never ending discussion in the literature: How can we estimate the predictive ability of multivariate models (and in particular QSAR models)? Here we cannot recapitulate the entire story, just refer to some basic sources. Generally, there are principally two different ways to evaluate the “goodness” of a QSAR model: to assess the model’s performance with regards to (i) description (fitting or recall), i.e. evaluating the performance on the existing data; and (ii) prediction, i.e. evaluating the performance on future data, also called external validation (how reliable a prediction can be made from the model for external data, such as for new molecules).

External validation is usually modelled by a single split (hold-out sample) in the belief that future compounds (objects, samples) will be derived from the same property distribution, which is more or less true for QSAR models within the applicability domain. If the future compounds diverge from the property distribution of the earlier ones (on which the model was built), then the model cannot be applied anymore without updating.

A common choice is to estimate the predictive ability using cross-validation; however, it is debatable how well cross-validation can mimic the prediction performance. Cross-validation is probably the most widely used method for estimating prediction error, but its various implementations inherently call for a compromise in terms of the bias-variance trade-off. As Hastie, Tibshirani and Friedman point out, “[...] five- or ten-fold cross-validation will overestimate the prediction error. Whether this bias is a drawback in practice depends on the objective. On the other hand leave-one-out cross-validation has low bias, but can have high variance. Overall, five- or tenfold cross-validation are recommended as a good compromise” (Hastie et al. 2009).

Some chemists also advocate a separation of an external part for testing (Esbensen and Geladi 2010), while others maintain the opposite: “hold-out sample is far inferior [as compared to leave-one-out cross-validation]” (Hawkins et al. 2003) or “hold-out samples are downward biased.[...] small independent hold-out samples are all but worthless” (Hawkins 2004).

As the machine learning community provides a plethora of novel techniques, which can produce 100% classification or error-free regression on the training set, the assessment of the predictive performance on future samples (i.e. validation, test) has gained increasing importance. There is no single best way to determine the predictive performance of a model, though some options such as leave-one-out cross-validation have become a kind of standard. We should emphasize the statistician’s view: “If possible, an independent sample should be obtained to test the adequacy of the prediction equation. Alternatively, the data set may be divided into three parts; one part to be used for model selection [model building or variable selection], the second part for the calibration of parameters in the chosen model and the last part for testing the adequacy of predictions” (Miller 1990).

In the machine learning field (artificial neural networks, support vector machines, etc.) this is the standard or at least the advocated practice. In many cases the insufficient number of samples leads to the division of the data into two parts. If the calibration of parameters is done using the same part of the data, substantial biases arise.

We should mention two recent sources with opposite conclusions: Esbensen and Geladi insist categorically on external validation with new measurements (Esbensen and Geladi 2010). Meanwhile, Gütlein et al. maintain that “contrary to current conception in the community, cross-validation may play a significant role in evaluating the predictivity of (Q)SAR models” (Gütlein et al. 2013). A somewhat intermediate opinion is presented by Gramatica, who agrees that cross-validation will generally give better and less variable results in terms of the prediction error for the available and modeled data, but also argues that only an additional “external evaluation” on totally new chemicals can represent a future working situation of the model (and thus, assess its predictivity) (Gramatica 2014). Her paper, together with an earlier work of her research group (Gramatica et al. 2012), also presents a thorough data splitting approach for external validation.

Recently, we have shown how one can identify the best (most consistent) performance indicators (merits) and demonstrated the capabilities of sum of ranking differences (SRD) in model selection and in the ranking of the performance merits. Based on two case studies from the literature (using a total of four training-test splits for the two case studies), we established that many of the performance parameters—if not all—for external validation are substantially inferior to other merits even if their application can be advantageous in some cases of data fusion (Rácz et al. 2015).

This work complements our earlier study on model performance parameters with two more case studies from the literature: a QSPR study employing a non-conventional technique, multivariate image analysis (MIA) to predict bioactivity-related properties of small peptides against Dengue virus 2 NS3 proteases (Silla et al. 2011), and a recent work by Roy et al. suggesting the use of error measures for QSAR model validation (Roy et al. 2016).

2 Model Performance Parameters (Merits)

Multivariate models can be evaluated with a large number of performance parameters (merits), including correlation-based (e.g. R^2 , Q^2) and error-like (e.g. MAE, RMSE) merits. In the QSAR modeling field—to the best of our knowledge—the QSARINS modeling software from the group of Paola Gramatica provides the largest pool of model performance parameters during QSAR modeling (Gramatica et al. 2013). A comprehensive summary of this set of performance parameters is available in our recent work (Rácz et al. 2015). In Table 1, the performance parameters occurring in at least one of the discussed case studies are included.

Table 1 Definition and description of the performance parameters compared

Performance parameter	Calculated during ^a	Formula ^b	Description
R^2, R_{ext}^2	Training, external validation	$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS}$	Explained variance; coefficient of determination, square of the multiple correlation coefficient
$RMSE$	Training, internal and external validation	$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$	Root mean square error
MAE	Training, internal and external validation	$MAE = \frac{\sum_{i=1}^n y_i - \hat{y}_i }{n}$	Mean absolute error
CCC	Training, internal and external validation	$CCC = \frac{2 \sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^n (y_i - \bar{y})^2 + \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 + n(\bar{y} - \bar{\hat{y}})^2}$	Coefficient of concordance, concordance correlation coefficient (Lin 1989, 1992)
$PRESS$	Internal, external validation	$PRESS = \sum_{i=1}^n (y_i - \hat{y}_{i/i})^2$	Predicted residual sum of squares (either cross-validated or calculated on the external set)
Q_{LOO}^2	Internal validation	$Q_{LOO}^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_{i/i})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{PRESS}{TSS}$	Leave-one-out cross-validated square of the (multiple) correlation coefficient
Q_{F1}^2	External validation	$Q_{F1}^2 = 1 - \frac{\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{ext}} (y_i - \bar{y}_{TR})^2}$	Definition 1 in Ref. (Consonni et al. 2010) for Q^2 of the external test set (Schüürmann et al. 2008), TR: training set
Q_{F2}^2	External validation	$Q_{F2}^2 = 1 - \frac{\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{ext}} (y_i - \bar{y}_{EXT})^2}$	Definition 2 in Ref. (Consonni et al. 2010) for Q^2 of the external test set (Shi et al. 2001), EXT: external test set

^aParameters that are calculated for more than one subsets are indexed in the main text: CV for cross-validation, EXT for external validation

^bThe following notation is used: y_i single experimental value; \bar{y} mean of experimental values; \hat{y}_i single predicted value; $\bar{\hat{y}}$ mean of predicted values; $\hat{y}_{i/i}$ predicted value for the i th sample when the i th sample is left out from the training; n number of samples; i sample index

2.1 Case Study 1

The MIA-QSPR application of Silla et al. (2011) involves the comparison of the correlation coefficient R^2 and the root mean square error RMSE, for calibration (cal), leave-one-out cross-validation (loo) and external validation (ext); a total of six performance parameters. While the selection is of moderate size, it provides an illustrative, balanced distribution of performance merits (two for calibration, two for cross-validation and two for external validation) to be compared. (Nonetheless, our recent work has shown that the outcome of SRD calculations is not—or only negligibly—influenced by the apparent “overweighting” of some methods (Bajusz et al. 2015).)

2.2 Case Study 2

In contrast, the article of Roy et al. (2016) on QSAR model validation deals exclusively with external validation merits. It is also interesting to know, which external merit(s) is (are) acceptable, preferable or which one(s) should be avoided. This work originally reports eight performance parameters for numerous QSAR models, and was complemented with PRESS values from the courtesy of Prof. Kunal Roy, arriving at a total of ten performance parameters. (PRESS values—along with multiple other merits—were calculated for the whole dataset, as well as for 95% of the data points, after omitting 5% high residual data points.)

3 Data Preprocessing Methods

Performance parameters can be distributed into two groups, which are scaled reversely: similarity (correlation) coefficient-like and error-like measures. To obtain comparable results we reversed the scaling of the error-like measures. Some well-known data preprocessing methods were used for the datasets: normalization (to unit length), rank transformation, range scaling, and standardization. The techniques are discussed in details below.

3.1 Normalization (NOR)

Normalization has several types, such as unit vector, area and mean normalization. Normalization based on area is used mostly in chromatography or spectroscopy, because it means that the observations are divided with the sum of all peak area. Mean normalization can be considered a classic choice: here the observations are

divided with the row average. Unit vector normalization is also popular, as it is frequently used in the preprocessing phase of pattern recognition methods. In our case, the latter was used. Its basic idea is that all variables are scaled to unit length, which means that the elements of a column are divided with their Euclidian distances of each column:

$$x_{i,j}^{normalized} = \frac{x_{i,j}}{\sqrt{\sum x_j^2}}, \quad (1)$$

Here j means the running index of columns.

3.2 Rank Transformation (RNK)

Rank transformation is the simplest data transformation technique, because in this case the only task is to order the values of a column (variable) in increasing (or in the reverse case: decreasing) magnitude and give a rank number to each value in the column. Thus the scale of the values will be between zero and the number of samples.

3.3 Range Scaling (SCL)

With the use of range scaling the variables are transformed into the [0; 1] (or other pre-defined) interval in a simple way:

$$x_{i,j}^{range\ scaled} = \frac{(x_{i,j} - Min(x_j))}{(Max(x_j) - Min(x_j))}, \quad (2)$$

where j means the running index for columns: 1, 2, ..., m. In this case there will be at least one zero and one unity in the dataset (or in each column, in case of more variables) by definition. Range scaling is very sensitive to outliers. Both range scaling and standardization increase the measurement errors. Range scaled values can be easily inverted:

$$x_{i,j}^{reversely\ scaled} = 1 - \frac{(x_{i,j} - Min(x_j))}{(Max(x_j) - Min(x_j))} \quad (3)$$

3.4 Standardization (STD)

In the case of standardization the centered matrix is divided with the column standard deviations. It is absolutely necessary, if the variables in the dataset are expressed in different units. Standardization can transform the variables to the same scale. In this way the variables are scaled to unit standard deviation. Standardization can be used in two different forms: row-wise and column-wise. While row-wise standardization is more important in the field of spectroscopy, in our case the column-wise version was used. The equation of the standardization process is the following:

$$x_{i,j}(\text{standardized}) = \frac{x_{i,j} - \text{Average}(x_j)}{\text{standard deviation}(x_j)} \quad (4)$$

4 Sum of Ranking Differences (SRD)

Sum of (absolute) ranking of differences is a novel and general ranking (ordering) and pattern recognition method for the comparison of methods, models and other types of features (variables) (Héberger 2010; Kollár-Hunek and Héberger 2013).

In the beginning the dataset should be compiled in the following format: the variables are arranged in the columns and the samples (observations, compounds) are in the rows. A reference column is also needed for the calculation, which can contain exact reference values, but row average, minimum or maximum values are also applicable as consensus approaches. (The choice depends on the dataset, e.g., minimums for error rates and maximums for non-error rates are suitable choices.)

In the first step the compounds (samples, observations) are ranked in every column (in the reference column, as well) in increasing magnitude. In the following step, differences are calculated between the ranks of the reference values and the ranks of each variable, for each row (sample). Finally the (absolute) differences are summed in every column: these are the SRD values, based on which the different models and methods can be compared. The smaller the SRD value, the better the method (more consistent with the reference), thus the best features are close to zero.

The validation of SRD calculations is carried out with a randomization test and a bootstrap-like cross-validation. (If the number of cases is smaller than fourteen, leave-one-out cross-validation is used.)

The final result of SRD is an ordering of methods (models, features, etc.), visualized on a plot, where both the x and the (left) y axis show the same SRD values. (Thus, the SRD values are lines instead of points in the plot.) The information is carried by the location of the lines and their proximity to each other and not by the height of the lines. Additionally, a Gauss-like curve corresponding to the distribution of SRD values of the randomization test is plotted, with frequency values on the right y axis. Features that overlap with the 95% of the Gauss-like

curve are not significantly better than the use of random numbers in terms of their ranking behavior, as compared to the reference (the 5% error limit is marked with dotted lines and abbreviation of XX1 in the SRD plots: anything below this limit is significantly different from random ranking at the 5% error level. Similarly, XX19 means 95% confidence, Med is the abbreviation for median).

The results of cross-validated SRD are favorably presented in Box and Whisker plots and can constitute the input of factorial ANOVA analysis in those cases, where there is more than one factor (indicator or grouping variable) present. The basic idea of ANOVA is that it tests the significance of differences between the group averages (where samples are grouped according to the indicator variables). ANOVA is a parametric technique and assumes (multi)normal distribution. In the case of factorial ANOVA we can use more than one factor, which means that we can test all the group averages with different group systems one by one and together as well.

5 Analysis of Variance (ANOVA)

ANOVA is a technique used to assess effects of the categorical factors and their interactions (Lindman 1991). The following model was considered:

$$\text{SRD} = b_0 + b_1 * I1 + b_2 * I2 + b_{12} * I1 * I2 \quad (5)$$

where SRD stands for the sum of absolute ranking differences, $I1$ is the type of preprocessing (four levels NOR, RNK, SCL, STD), $I2$ is the performance parameter: 6 levels in *Case study 1* (RMSE_{cal} , RMSE_{ext} , RMSE_{100} , r_{cal}^2 , r_{ext}^2 , r_{100}^2) and 10 levels in *Case study 2* (CCC , PRESS_{95} , PRESS_{100} , MAE_{95} , MAE_{100} , R_{100}^2 , Q_{F1-95}^2 , Q_{F2-95}^2 , Q_{F1-100}^2 , Q_{F2-100}^2).

Seven repetitions allow us to test the significance of factors and their interactions.

Variance analysis decomposes the effect of the different factors on the SRD values. This unique combination of SRD and ANOVA provides not only the relative importance of factors, but also an overall evaluation, and has proven to be successful in earlier cases, such as comparing evaluation techniques for genotoxicity measurements (Héberger et al. 2014) and comparing similarity measures for molecular fingerprints (Bajusz et al. 2015).

SRD analyses have been carried out with our own scripts, including the recently published SRD-COVAT heatmaps (Andrić et al. 2016), all of which are downloadable from our website: <http://aki.ttk.mta.hu/srd/>.

ANOVA calculations have been carried out with STATISTICA (version 12.5, StatSoft, Inc., Tulsa, OK 74104, USA, 2014).

6 Results and Discussion

Similarly to our earlier paper (Rácz et al. 2015) we have chosen two examples from the literature as case studies for the comparison of various model performance parameters applied in the QSAR modeling field. While previously we have taken only the raw data from the publications and carried out QSAR modeling ourselves, this time we have selected two papers that have reported a selection of performance parameters for several models that were compared by the authors.

6.1 Case Study 1

In a 2011 study Silla et al. have applied multivariate image analysis of 2D chemical structures to develop QSPR models for the prediction of bioactivity-related properties (substrate cleavage rate constant k_{cat} and Michaelis constant K_M) of small peptides against Dengue virus 2 NS3 proteases (Silla et al. 2011). Since image analysis is an inherently high-dimensional task (each pixel of the image can be considered a dimension), a suitable variable selection technique is of paramount importance in such studies. To that end, the authors compared numerous PLS models, where the variables were selected with one of (or a combination of) three variable selection methods: interval PLS (iPLS), genetic algorithm (GA) and ordered predictors selection (OPS).

Table 2 of the mentioned paper summarizes six performance parameters for 22 models, namely R^2 and RMSE values for calibration, leave-one-out cross-validation and external validation (see Table 1 for definitions). The external test set was compiled randomly and contained 11 compounds (vs. the 43 compounds in the training set). The data in the mentioned table are suitable for a detailed statistical analysis, for a fair comparison of performance parameters (merits).

As the merits are measured on different scales, first they have to be placed on the same scale. Four possibilities have been selected for this task: normalization, rank transformation, range scaling and standardization. (Naturally the error-like

Table 2 Results of two-way ANOVA conducted on the cross-validated SRD values, with the data preprocessing methods ($I1$) and the performance parameters ($I2$) as indicator variables, for Case study 1

	SS	DOF	MS	F	p
Intercept	231607.4	1	231607.4	81233.46	0.000000
$I1$	8.6	3	2.9	0.03	0.991579
$I2$	29989.8	5	5998.0	69.61	0.000000
$I1 * I2$	1292.5	15	86.2	6.08	0.000000
Error	2380.7	168	14.2		

SS sum of squares, DOF degrees of freedom, MS mean squares

measures should be reversed to obtain comparable quantities.) Thus, four 6×22 input matrices were formed according to the data preprocessing techniques.

During the SRD analysis, average was used as the benchmark (reference column) with the consideration that all performance parameters express some prediction ability with error. (The maximum likelihood principle would suggest the usage of average as the best estimation.) Figure 1 shows the ordering result of the SRD procedure on the standardized dataset. Here, R^2 values based on calibration and cross-validation are the two performance parameters that are most consistent with the reference, while $RMSE_{ext}$ is over the 5% limit (i.e. indistinguishable from random ranking). The process was repeated for all the four data preprocessing methods and all of the four matrices were subjected to a sevenfold cross-validation. In such a way, 192 SRD values were calculated showing characteristic patterns according to the factors: performance merits (Fig. 2) and data preprocessing techniques (Fig. 3). As an additional validation step, we have made sure that the SRD values resulting from the whole dataset are in conformity with the SRD value distribution acquired from sevenfold cross-validation (data not shown).

While there is a generally good agreement between ranking, range scaling and standardization, normalization to unit length is peculiar, sometimes the worst (largest), sometimes the best (smallest) one among the preprocessing methods. However, the differences among performance parameters cannot be traced back to the choice of different pretreatment methods, as demonstrated by the ANOVA results in Table 1.

Two-way analysis of variance (ANOVA) has been carried out on the SRD results, with the data preprocessing methods (I1) and the performance merits (I2) as

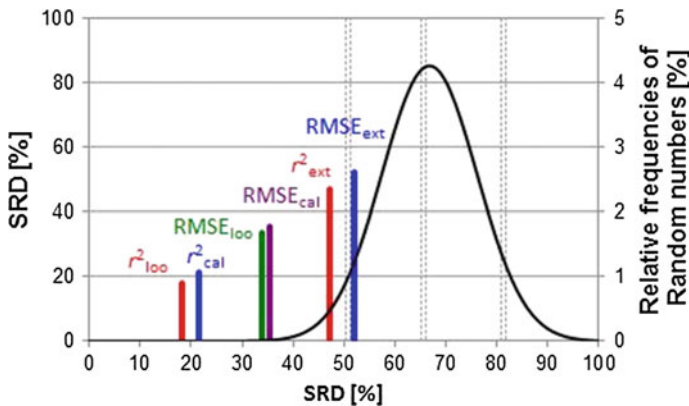


Fig. 1 Scaled SRD values (between 0 and 100) compared to random ranking (*black* Gaussian curve) for the standardized dataset. In this example, $RMSE_{ext}$ overlaps with the Gaussian curve and is thus not significantly different from random ranking. r_{100}^2 is the most consistent with the reference (in terms of ranking the models)

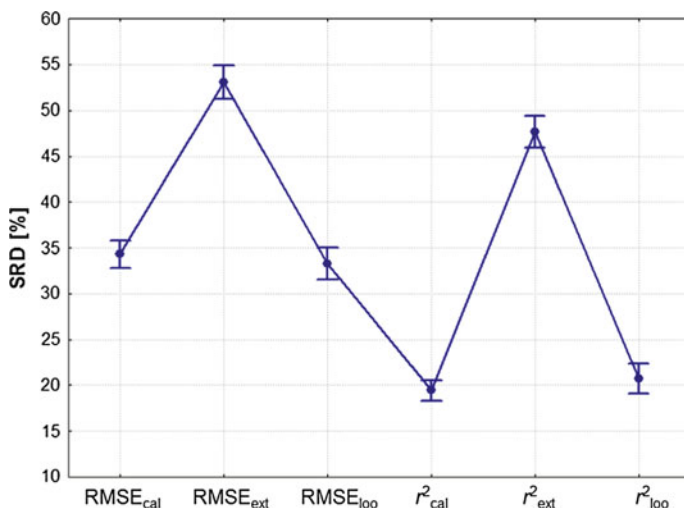


Fig. 2 Sevenfold cross-validated SRD results for the comparison of performance parameters. r^2 values based on calibration and leave-one-out cross-validation are the most consistent metrics (as they display the smallest SRD values), RMSE values from the same procedures are intermediate and the r^2 and RMSE values based on external validation are the least consistent with the reference (average)

the two indicator variables. Based on the ANOVA analysis we conclude that the choice of the data preprocessing method does not influence the SRD values significantly (at a 5% error level suggesting that no “artificial” effect was introduced with data preprocessing), while the choice of the performance merit as well as the combination of the two factors does.

6.2 Case Study 2

A 2016 study by Roy et al. promotes the use of error measures for the evaluation of QSAR models, as a more advantageous alternative to “classic” correlation-based metrics (Roy et al. 2016). The authors argue that while R^2 -based performance parameters are easier to comprehend (due to their fixed [0; 1] range), they are highly dependent on the range of the response values. However, the study deals exclusively with external validation parameters. In addition, a guideline is proposed to assess the quality of predictions based on the mean absolute error (MAE) and its standard deviation computed from 95% of the test set predictions (after omitting 5% high residual data points). Tables 1, 2 and 3 of Ref. [11] report various performance parameters based on the external validation of an abundance of QSAR models, and have formed the basis of our analysis. The original tables were complemented with PRESS values from the courtesy of Prof. Kunal Roy. Interestingly, Q_{F3}^2 has been

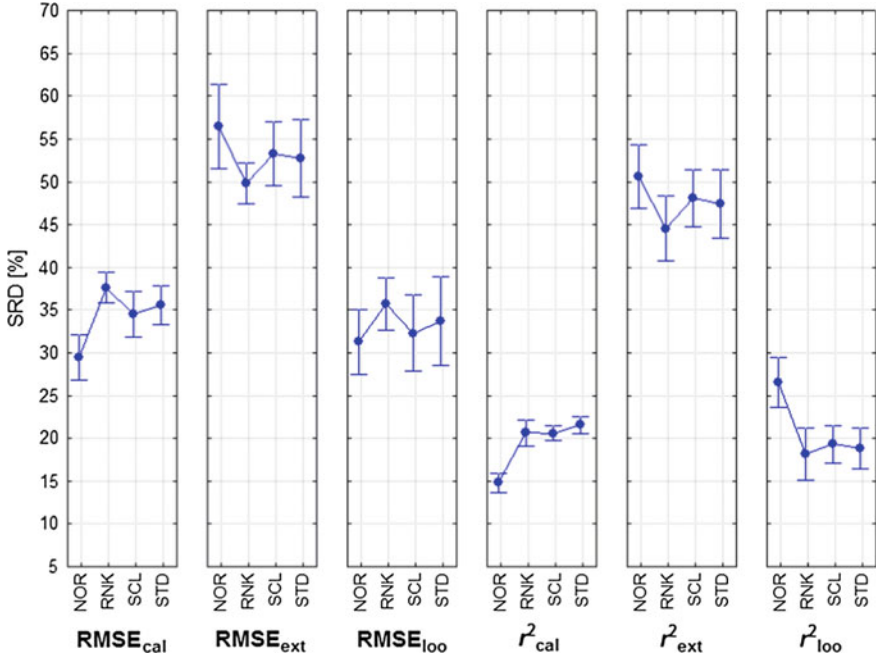


Fig. 3 Effects of preprocessing to the SRD values of the performance merits. The preprocessing techniques are generally in good agreement, with the exception of normalization in some cases (such as for r_{cal}^2 and r_{loo}^2)

left out from the evaluation, though Consonni et al. suggested its superiority (Consonni et al. 2010). On another note, Chirico and Gramatica suggested the favorable usage of the coefficient of concordance (Chirico and Gramatica 2011).

The same SRD procedure has similarly been carried out as for *Case study 1*, the average was used as the reference here, as well (see Fig. 4). All of the four data preprocessing methods were applied as in *Case study 1*: standardization, normalization, range scaling and rank transformation. Analysis of variance has confirmed the conclusions drawn in the first case study: the choice of the data preprocessing method is not a significant factor (see Table 2) suggesting that no “artificial” effect was introduced with data preprocessing.

One can argue that the SRD results are principally determined by the selection of the reference (benchmark) column, which is true to some extent (but overlooks the maximum likelihood principle and the superiority of the consensus approach over an individual reference variable). Therefore, we have elaborated a technique to examine the underlying data structure to a finer “resolution”. In this case, each variable (column) is used as the reference, one at a time and a color-coded matrix (heatmap) is compiled from the results. This approach was termed COVAT–Comparison with One Variable at a Time—and was introduced in our recent paper

Table 3 Results of two-way ANOVA conducted on the cross-validated SRD values, with the data preprocessing methods ($I1$) and the performance parameters ($I2$) as indicator variables, for *Case study 2*

	SS	DOF	MS	F	p
Intercept	589413.5	1	589413.5	88787.07	0.000000
$I1$	34.6	3	11.5	1.74	0.159475
$I2$	45074.0	9	5008.2	754.42	0.000000
$I1 * I2$	884.6	27	32.8	4.94	0.000000
Error	1858.8	280	6.6		

SS sum of squares, DOF degrees of freedom, MS mean squares

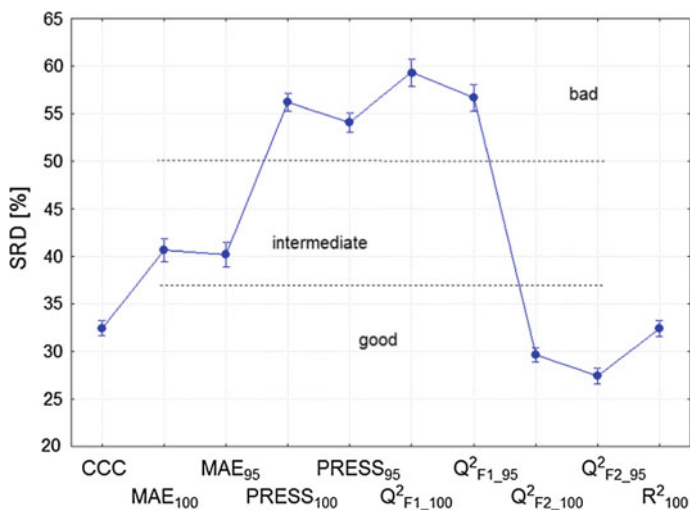


Fig. 4 Sevenfold cross-validated SRD results for the comparison of external validation parameters (with average values as the reference vector). It is relatively easy to classify these performance parameters into good (consistent), intermediate and bad (least consistent) ones considering the SRD gaps between groups. The fact that the concordance correlation coefficient is among the good merits strengthens Chirico and Gramatica's recommendation (Chirico and Gramatica 2011) about its usefulness

on lipophilicity scales (Andrić et al. 2016). The different benchmarks eliminate the problem of golden standard selection: the grouping of the SRD values obtained with the various reference vectors can reveal the underlying connections between the examined variables (here, performance parameters).

For the heatmap calculations, standardization has been selected as the data preprocessing method—keeping in mind that the factor of data preprocessing was proven to be *not* significant. The results are shown in Fig. 5.

Figure 5a highlights three groups of external validation merits: group 1 (upper left part) contains $Q^2_{F2_95}$, $Q^2_{F2_100}$, R^2_{100} and CCC, group 2 (middle) contains MAE₉₅ and

(a)	Colors:	x <=	10	30	50	70	90				
			20	40	60	80	100				
	$Q^2_{F2,95}$	$Q^2_{F2,100}$	R^2_{100}	CCC	MAE ₉₅	MAE ₁₀₀	$Q^2_{F1,95}$	$Q^2_{F1,100}$	PRESS ₉₅	PRESS ₁₀₀	
$Q^2_{F2,95}$	0.0	5.6	17.6	10.7	50.6	50.4	60.9	62.9	74.8	76.9	
$Q^2_{F2,100}$	5.6	0.0	17.6	9.4	53.6	53.0	62.7	63.0	77.6	79.7	
R^2_{100}	17.4	17.6	0.0	11.1	59.4	60.2	54.1	55.1	72.8	75.3	
CCC	10.6	9.4	11.2	0.0	55.5	55.8	61.8	63.0	79.0	81.1	
MAE ₉₅	50.6	53.6	59.4	55.5	0.0	5.1	74.0	76.6	44.4	47.1	
MAE ₁₀₀	50.4	53.0	60.2	55.8	5.1	0.0	75.4	77.6	46.1	48.8	
$Q^2_{F1,95}$	61.0	62.7	54.2	61.8	74.0	75.4	0.0	10.0	56.8	57.6	
$Q^2_{F1,100}$	62.9	63.0	55.1	62.9	76.6	77.6	9.9	0.0	58.2	57.8	
PRESS ₉₅	74.8	77.6	72.8	79.0	44.4	46.1	56.7	58.2	0.0	5.4	
PRESS ₁₀₀	76.9	79.7	75.3	81.1	47.1	48.8	57.5	57.8	5.4	0.0	

(b)	Colors:	x <	57.6	<= x <	66.1	<= x <	74.8	<= x			
			XX1		Med		XX19				
	$Q^2_{F2,95}$	$Q^2_{F2,100}$	R^2_{100}	CCC	MAE ₉₅	MAE ₁₀₀	$Q^2_{F1,95}$	$Q^2_{F1,100}$	PRESS ₉₅	PRESS ₁₀₀	
$Q^2_{F2,95}$	0.0	5.6	17.6	10.7	50.6	50.4	60.9	62.9	74.8	76.9	
$Q^2_{F2,100}$	5.6	0.0	17.6	9.4	53.6	53.0	62.7	63.0	77.6	79.7	
R^2_{100}	17.4	17.6	0.0	11.1	59.4	60.2	54.1	55.1	72.8	75.3	
CCC	10.6	9.4	11.2	0.0	55.5	55.8	61.8	63.0	79.0	81.1	
MAE ₉₅	50.6	53.6	59.4	55.5	0.0	5.1	74.0	76.6	44.4	47.1	
MAE ₁₀₀	50.4	53.0	60.2	55.8	5.1	0.0	75.4	77.6	46.1	48.8	
$Q^2_{F1,95}$	61.0	62.7	54.2	61.8	74.0	75.4	0.0	10.0	56.8	57.6	
$Q^2_{F1,100}$	62.9	63.0	55.1	62.9	76.6	77.6	9.9	0.0	58.2	57.8	
PRESS ₉₅	74.8	77.6	72.8	79.0	44.4	46.1	56.7	58.2	0.0	5.4	
PRESS ₁₀₀	76.9	79.7	75.3	81.1	47.1	48.8	57.5	57.8	5.4	0.0	

Fig. 5 SRD-COVAT heatmaps of the external validation parameters in *Case study 2* with an equidistant (a) and a “Gaussian” (b) color coding. (Color references are provided on the upper parts of the images.) While *panel A* highlights four clusters of similar performance parameters, *panel B* provides information on the significance of SRD values, i.e. relative to the distribution of random rankings

MAE₁₀₀, group 3 (lower right) contains $Q^2_{F1,95}$, and $Q^2_{F1,100}$, PRESS₉₅, and PRESS₁₀₀. While the first two groups confirm the conclusions based on Fig. 4 completely, the third “group” can be further divided based on Q^2_{F1} and PRESS values, though they have similar (sometimes overlapping) SRD distributions against the average as reference (see Fig. 4). Additionally, the pairs of performance parameters calculated from the whole dataset and 95% are close to each other, as expected.

Figure 5b offers even more intriguing results, as it shows the SRD values relative to the SRD distribution of random rankings (consult the Gaussian curve on Fig. 1 for reference): cells of any other color than white denote that there is no correspondence between the rankings produced by the two external validation parameters indicated in the implied row and column headers. As there are many such cells in the table, we can conclude that the ranking results obtained by most of these (external) performance merits are highly divergent. Ultimately, this can safely

be considered as a conclusion supporting the preference of performance parameters based on cross-validation, since there seems to be little consensus among those based on external validation (see Fig. 5). From the external validation metrics, we would suggest the use of Q_{F2}^2 , R_{100}^2 and CCC as the most consensual ones.

7 Conclusion

We have carried out a comparison of QSAR model performance parameters based on two case studies, with the combination of sum of ranking differences (SRD) and analysis of variance (ANOVA). The first case study has shown cross-validation based performance metrics to be more consistent with the consensus ranking than those based on external validation. In the second case study, we have compared some members of the latter group in more detail and have shown that the rankings produced by them are greatly divergent. The results presented here corroborate our earlier, recently published findings (Rácz et al. 2015) on diverse data sets of independent literature sources.

Showing a model to be predictive for a small external test set does not necessarily mean that it will be predictive for molecules outside of this test set. In other words, in the case of external validation we are delivered to a random test, which might be informative but not necessarily. While we agree that a more meticulous training-test splitting approach (such as the one presented by Gramatica et al. (2012)) can significantly improve the reliability of external validation, we would still advise against overemphasizing model performance parameters based on external validation, or preferring them over the ones derived from cross-validation. (In our opinions, a consensus approach might be the best choice here.) In the lack of sufficient test data (which is often the case in QSAR modeling), our results reinforce the conclusions of Hawkins et al. (2003), who advise against small holdout samples (to avoid the loss of information in model building) and recommend cross-validation instead.

Acknowledgement The work was supported by the Hungarian Scientific Research Fund (OTKA, grant number K 119269).

References

- Andrić, F., Bajusz, D., Rácz, A., et al. (2016). Multivariate assessment of lipophilicity scales—Computational and reversed phase thin-layer chromatographic indices. *Journal of Pharmaceutical and Biomedical Analysis*, 127, 81–93. doi:10.1016/j.jpba.2016.04.001.
- Bajusz, D., Rácz, A., & Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7, 20. doi:10.1186/s13321-015-0069-3.
- Chirico, N., & Gramatica, P. (2011). Real external predictivity of QSAR models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance

- correlation coefficient. *Journal of Chemical Information and Modeling*, *51*, 2320–2335. doi:10.1021/ci200211n.
- Consonni, V., Ballabio, D., & Todeschini, R. (2010). Evaluation of model predictive ability by external validation techniques. *Journal of Chemometrics*, *24*, 194–201. doi:10.1002/cem.1290.
- Esbensen, K. H., & Geladi, P. (2010). Principles of proper validation: Use and abuse of re-sampling for validation. *Journal of Chemometrics*, *24*, 168–187. doi:10.1002/cem.1310.
- Gramatica, P. (2014). External evaluation of QSAR models, in addition to cross-validation: Verification of predictive capability on totally new chemicals. *Molecular Informatics*, *33*, 311–314. doi:10.1002/minf.201400030.
- Gramatica, P., Cassani, S., Roy, P. P., et al. (2012). QSAR Modeling is not “push a button and find a correlation”: A case study of toxicity of (Benzo-)triazoles on Algae. *Molecular Informatics*, *31*, 817–835. doi:10.1002/minf.201200075.
- Gramatica, P., Chirico, N., Papa, E., et al. (2013). QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. *Journal of Computational Chemistry*, *34*, 2121–2132. doi:10.1002/jcc.23361.
- Gütlein, M., Helma, C., Karwath, A., & Kramer, S. (2013). A large-scale empirical evaluation of cross-validation and external test set validation in (Q)SAR. *Molecular Informatics*, *32*, 516–528. doi:10.1002/minf.201200134.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). Cross-Validation. *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed., pp. 241–249). New York: Springer.
- Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, *44*, 1–12. doi:10.1021/ci0342472.
- Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences*, *43*, 579–586. doi:10.1021/ci025626i.
- Héberger, K. (2010). Sum of ranking differences compares methods or models fairly. *TrAC Trends in Analytical Chemistry*, *29*, 101–109.
- Héberger, K., Kolarević, S., Kračun-Kolarević, M., et al. (2014). Evaluation of single-cell gel electrophoresis data: Combination of variance analysis with sum of ranking differences. *Mutation Research, Genetic Toxicology and Environmental Mutagenesis*, *771*, 15–22. doi:10.1016/j.mrgentox.2014.04.028.
- Kollár-Hunek, K., & Héberger, K. (2013). Method and model comparison by sum of ranking differences in cases of repeated observations (ties). *Chemometrics and Intelligent Laboratory Systems*, *127*, 139–146. doi:10.1016/j.chemolab.2013.06.007.
- Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, *45*, 255–268.
- Lin, L. I.-K. (1992). Assay validation using the concordance correlation coefficient. *Biometrics*, *48*, 599. doi:10.2307/2532314.
- Lindman, H. R. (1991). *Analysis of variance in experimental design*. New York: Springer.
- Miller, A. (1990). *Subset selection in regression*. London: Chapman and Hall.
- Rácz, A., Bajusz, D., & Héberger, K. (2015). Consistency of QSAR models: Correct split of training and test sets, ranking of models and performance parameters. *SAR and QSAR in Environmental Research*, *26*, 683–700. doi:10.1080/1062936X.2015.1084647.
- Roy, K., Das, R. N., Ambure, P., & Aher, R. B. (2016). Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemometrics and Intelligent Laboratory Systems*, *152*, 18–33. doi:10.1016/j.chemolab.2016.01.008.
- Schüürmann, G., Ebert, R.-U., Chen, J., et al. (2008). External validation and prediction employing the predictive squared correlation coefficient test set activity mean vs training set activity mean. *Journal of Chemical Information and Modeling*, *48*, 2140–2145. doi:10.1021/ci800253u.
- Shi, L. M., Fang, H., Tong, W., et al. (2001). QSAR models using a large diverse set of estrogens. *Journal of Chemical Information and Modeling*, *41*, 186–195. doi:10.1021/ci000066d.
- Silla, J. M., Nunes, C. A., Cormanich, R. A., et al. (2011). MIA-QSPR and effect of variable selection on the modeling of kinetic parameters related to activities of modified peptides against dengue type 2. *Chemometrics and Intelligent Laboratory Systems*, *108*, 146–149. doi:10.1016/j.chemolab.2011.06.009.

Part II

Methods

Structural, Physicochemical and Stereochemical Interpretation of QSAR Models Based on Simplex Representation of Molecular Structure

P. Polishchuk, E. Mokshyna, A. Kosinskaya, A. Muats, M. Kulinsky, O. Tinkov, L. Ognichenko, T. Khristova, A. Artemenko and V. Kuz'min

Abstract In this chapter we describe different structural, physicochemical and stereochemical approaches towards interpretation of QSAR models based on simplex representation of molecular structure (SiRMS). These techniques are feasible due to the flexible nature of SiRMS, which may encode not only structural and physicochemical features of molecules, but also stereochemical ones (to represent molecules with different types of chirality). The developed approaches to structural

P. Polishchuk (✉)

Faculty of Medicine and Dentistry, Institute of Molecular and Translational Medicine, Palacký University and University Hospital in Olomouc, Hněvotínská 1333/5, 779 00 Olomouc, Czech Republic
e-mail: pavlo.polishchuk@upol.cz

E. Mokshyna · A. Kosinskaya · M. Kulinsky · L. Ognichenko · T. Khristova · A. Artemenko · V. Kuz'min

A.V. Bogatsky Physico-Chemical Institute of National Academy of Sciences of Ukraine, Lustdorfskaya Doroga 86, Odessa 65080, Ukraine
e-mail: mokshinaelena@ukr.net

A. Kosinskaya
e-mail: nikang@ukr.net

M. Kulinsky
e-mail: docmax@inbox.ru

L. Ognichenko
e-mail: ogni@ukr.net

T. Khristova
e-mail: tanya07-88@mail.ru

A. Artemenko
e-mail: artanat@ukr.net

A. Muats
Chemical Department, Odessa National University, Dvoryanskaya 2, Odessa 65082, Ukraine
e-mail: nandorua92@gmail.com

and physicochemical interpretation do not depend on used machine learning methods that makes it possible to easily interpret traditional “black box” models like Support Vector Machine and Random Forest. We demonstrated an applicability of the developed interpretation approaches in a number of case studies including classical Hammett and Free-Wilson analysis, as well as several data sets with various physical and biological end-points. A good correspondence of the interpretation results with classical Hammett and Free-Wilson approaches supports validity of the proposed approaches. The analysis of different data sets with different end-points showed three possible scenarios of QSAR models’ interpretation depending on the mechanisms of action for studied compounds that brings us to a conclusion that despite all models are interpretable, not all end-points are. The stereochemical interpretation was applied to the classical Cramer’s set of steroids and to the data set that includes compounds with mixed central and axial chirality. In both cases we demonstrated the substantial contribution of the chiral descriptors in 2.XD QSAR models and revealed certain stereochemical features, which have the biggest contributions to investigated properties. As SiRMS represents an attractive framework for developing predictive and interpretable models, we developed several open-source software tools to make it available for the community. They are discussed at the end of the chapter.

Keywords QSAR model interpretation • Simplex representation of molecular structure • SiRMS • Structural and physicochemical interpretation of QSAR models • Stereochemical interpretation of QSAR models • Free-Wilson models • Hammett constants • Critical properties • Antagonists of fibrinogen receptor • Acute oral toxicity

1 Introduction

Since the introduction of the simple linear models, the modeling community shifted its attention to much more complex non-linear machine learning approaches (Support Vector Machine, Neural Network, Random Forest) as well as consensus modeling (when various individual models are combined to predict target properties). A better predictive ability of the latter caused such shift because they significantly outperform the linear models in many tasks. But interpretation of the

O. Tinkov

T. G. Shevchenko Transdnistria State University, 25 Oktyabrya 107, 3300 Tiraspol,
Transdnistria, Republic of Moldova
e-mail: tinkov84@mail.ru

V. Kuz'min (✉)

Odessa National Polytechnic University, Shevchenko Av., 1, Odessa 65044, Ukraine
e-mail: theorchem@gmail.com

non-linear and consensus models is less straightforward than of the linear models, which promotes the idea of a certain trade-off between predictivity and interpretability of QSAR models, so-called “two QSARs” (Guha 2008; Fujita and Winkler 2016).

Nevertheless, interpretation of QSAR models appears to be of fundamental importance (Cherkasov et al. 2014; Gasteiger 2016). The interpretation helps a researcher to understand established structure-property relationships. The interpretation results may be compared to empirical knowledge for a knowledge-based validation of QSAR models. If the interpretation results contradict experimentally observed structure-property relationships, this signifies that either a model is wrong or the empirical knowledge is incomplete. From an interpretation of QSAR models, a researcher can find desirable or undesirable structural motifs for a design of new compounds with improved properties. Interpretation provides information that not only helps to understand but also improves QSAR models.

Traditionally interpretation of QSAR models consists of two parts: interpretability of machine learning models and interpretability of descriptors. But recent advances made it possible to interpret any QSAR model regardless of the employed descriptors and/or machine learning approach (Polishchuk et al. 2013, 2016; Riniker and Landrum 2013; Sushko et al. 2014). In this chapter we present one of such universal approaches towards structural interpretation of QSAR models and its extensions, which allows a deeper analysis of captured structure-property relationships in physicochemical terms. The simplex representation of molecular structure (SiRMS) is a very powerful and flexible approach and it perfectly suits for structural and physicochemical interpretation of QSAR models. Within the SiRMS, one can also encode stereochemical information and, thus, perform stereochemical interpretation of QSAR models. Here, we demonstrate the applicability of simplex descriptors to development of interpretable models using several case studies.

2 Simplex Representation of Molecular Structure (SiRMS)

Throughout several last decades simplex representation of molecular structure was successfully applied to many different tasks (Kuz'min et al. 2005, 2008). Here, we describe some basics for understanding of interpretation results of QSAR models. The simplexes are tetraatomic fragments of fixed composition, topology, chirality and symmetry, whilst simplex descriptors are counts of the identical simplexes in a structure. The simplexes can be bound or connected (all atoms in a simplex are connected by bonds) or unbound or unconnected (one or more atoms are not connected to others in a simplex). The latter feature allows to encode structures consisting of separate fragments and/or stereochemistry of compounds. The SiRMS uses the labeling of atoms according to their physicochemical properties—partial atomic charges (representing electrostatic interactions), lipophilicity (hydrophobic

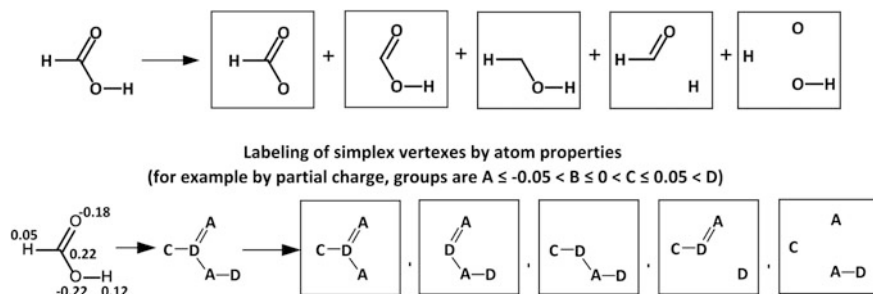


Fig. 1 2D simplex representation of molecular structure

Cahn-Ingold-Prelog configuration	S-isomer	R-isomer
Configuration based on simplexes	<u>RSSSS</u>	<u>RSSRR</u>

Fig. 2 An example of two compounds with similar configuration and their corresponding stereoconfiguration defined according to Cahn-Ingold-Prelog rules and simplex representation

interactions), polarizability (dispersive interactions), H-bond donor/acceptor (H-bonding), etc. When the values of atomic properties lay in a continuous scale (like partial atomic charges), the whole range of values is divided into a specific number of bins (usually 4–7), and each bin receives its own label. Such label serves as an atom label on the stage of simplex generation (Fig. 1). In this study, we have used simplex descriptors labeled by partial atomic charge, lipophilicity, refractivity and H-bonding. All the parameters were calculated using the Chemaxon cxcalc software tool (cxcalc).

The full set of chiral simplexes uniquely represents stereoisomers of any types (enantiomers or diastereomers with any chiral features: cis/trans bonds, chiral centers, etc.). Even for the simple molecules with one chiral center, SiRMS has certain advantages in comparison to classical Cahn-Ingold-Prelog (CIP) system as the molecules with similar configuration are represented as different stereoisomers (Fig. 2). Since CIP system solves only nomenclature problems, it sometimes prevents from distinguishing different classes of homochirality of chiral molecules.

Within the SiRMS, a chiral center is represented by the set of 5 simplexes. Canonical numbers are assigned to each atom in simplexes by known algorithms (Weininger et al. 1989). Thus, one can rank all simplexes according to the precedence of atoms in them (Fig. 3). The stereochemical representation of the

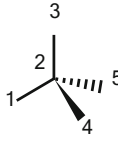
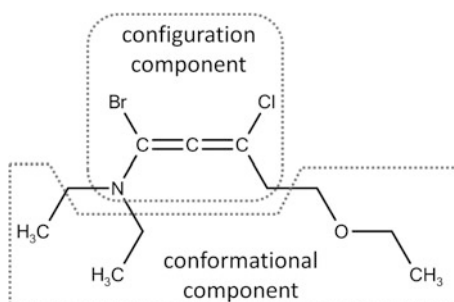
Atom ranks in simplexes	Simplex precedence	Simplex configuration	
1 2 3 4	1	R	 <p>RSRRR</p>
1 2 3 5	2	S	
1 2 4 5	3	R	
1 3 4 5	4	R	
2 3 4 5	5	R	

Fig. 3 An example of encoding of stereoconfiguration for a hypothetical molecule with one chiral center (numbers are canonical numbers of atoms obtained with conventional algorithms)

Fig. 4 Conformational and configuration components of 3D representation of molecules



compounds with a single chiral center is a sequence configuration of simplexes ordered by their precedence and for a hypothetical molecule depicted on Fig. 3 will be RSRRR. Obviously its enantiomers will have configuration SRSSS, all simplexes will have an opposite configuration.

If one calculates the configuration of molecules depicted in Fig. 2, three most precedent simplexes will have the same configuration that reflects common stereochemical features of molecules. A 3D structure of any molecule may be represented as a system of simplexes; and thus all the stereochemical peculiarities are taken into account. However, conformations of molecules, in which they bind to their functional targets, are usually not known and 3D approaches cannot be directly applicable.

For a 2D representation, several approaches exist that allow to encode stereoisomers (Golbraikh et al. 2001; Lukovits and Linert 2001; Aires-de-Sousa and Gasteiger 2002; Carbonell et al. 2013). But all of them are limited to the molecules with chiral centers only. The SiRMS allows to solve this task for molecules with different chiral features and, moreover, the conformation and configuration components of 3D representation can be separated. Thus, one may define configuration without concretizing a conformational part of 3D molecular representation (Fig. 4). To do this one needs to separate conformationally independent simplexes (whose configuration does not depend on conformation). These conformationally independent 3D simplex descriptors are concatenated to topological 2D simplex

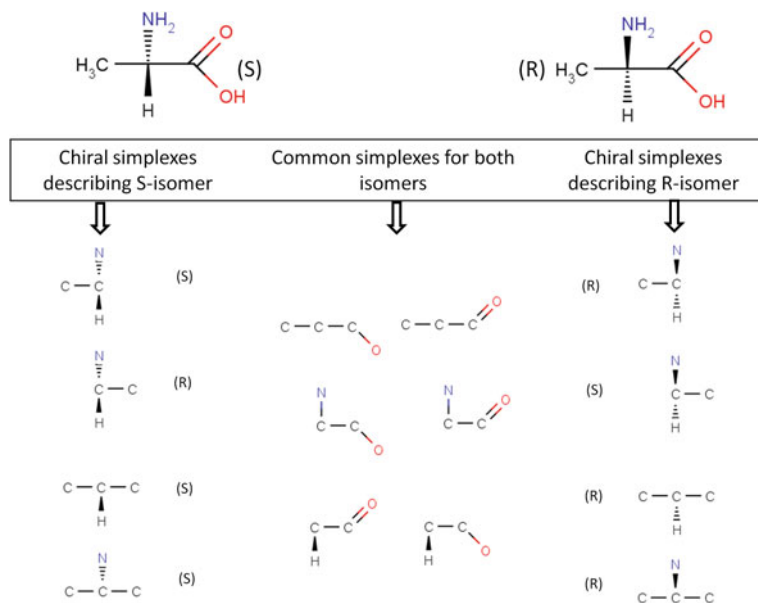


Fig. 5 An example of $(2 + 0.X)D$ simplex representation approach for enantiomers of alanine

descriptors. Such representation may be considered as intermediate between 2D and 3D representation— $(2 + 0.X)D$.

The $(2 + 0.X)D$ SiRMS approach on enantiomers of alanine returns 2D simplexes common for both enantiomers and conformationally independent 3D simplexes that discriminate those enantiomers (Fig. 5). Inherently, stereochemical interpretation estimates the 0.X value, in other words the contribution of the descriptors representing stereoconfiguration of molecules.

3 Structural and Physicochemical Interpretation of QSAR Models

3.1 Structural Interpretation

The idea of structural interpretation relates to the matched molecular pair approach and employs the assumption that the contribution of the fragment of interest (C) can be calculated as a difference between predicted property values for the initial molecule (A) and the counter-fragment (B) which remains after virtual removal of the fragment of interest from the initial molecule (Fig. 6). Thus, one may estimate contributions of any fragment or single atoms. This interpretation procedure does

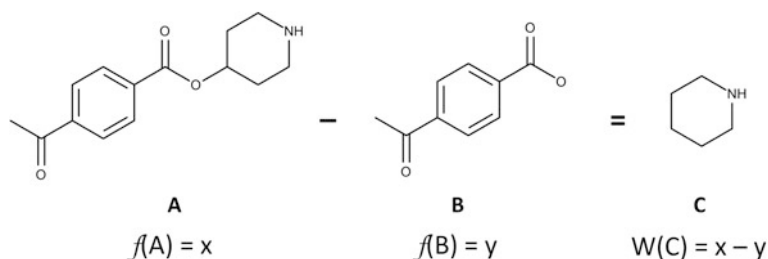


Fig. 6 An example of structural interpretation of QSAR models. f is a QSAR model which returns predicted property values for initial compound A and counter-fragment B which remains after virtual removal of the fragment of interest C. $W(C)$ is a contribution of the fragment C to the investigated property

not depend on the descriptors and the machine learning methods and can be considered as universal (Polishchuk et al. 2013).

The SiRMS representation perfectly fits the described approach as it allows to estimate contributions of any arbitrary atom subset (connected or disconnected), scaffolds and linkers. Due to the presence of the disconnected simplexes the removal of fragments (scaffolds or linkers) leading to the multiple disconnected parts can be easily performed.

3.2 Physicochemical Interpretation

Using the interpretation scheme described above one can estimate the contributions in different physicochemical terms. To perform physicochemical interpretation, the initial compound should be encoded by descriptors which represent certain physicochemical properties. The descriptors representing the same physicochemical property should be grouped together. First the fragment of interest (C) is removed from the initial compound (A). Then only descriptors belonging to the certain group are calculated for the counter-fragment (B) (Fig. 7). Thus fragment C is virtually removed in terms of the certain group of descriptors only. The calculated difference between predicted property values for the initial molecule (A) and the counter-fragment (B) will represent the contribution of the fragment C in terms of the selected group of descriptors and thus reflect contribution of a particular physicochemical property encoded by them.

Within the SiRMS, each molecule can be represented by simplexes labeled by certain physicochemical properties as described above (partial atomic charge, lipophilicity, H-bonding and refractivity). Thus the SiRMS perfectly suits for physicochemical interpretation of QSAR models (Polishchuk et al. 2016).

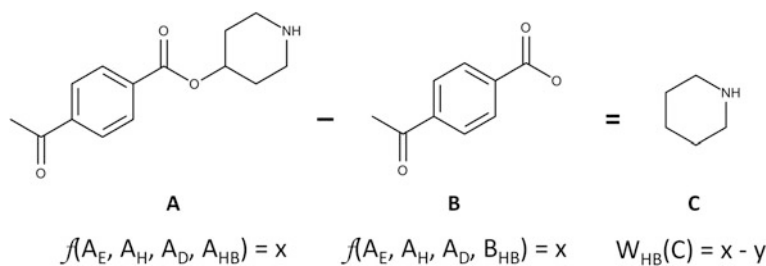


Fig. 7 An example of physicochemical interpretation of QSAR models. A_E , A_H , A_D , A_{HB} are groups of descriptors representing electrostatic, hydrophobic, dispersive and hydrogen bonding terms of the molecule A, correspondingly. B_{HB} is the group of descriptors representing hydrogen bonding of the counter-fragment B. $W_{HB}(C)$ is the contribution of the fragment C regarding the hydrogen bonding term

Table 1 Strategies of fragments selection and assembling

Scenario (data set type)	Do specific interactions exist and cannot be disregarded?	Is the position of a ligand towards its target known?	Fragments selection and grouping
1 (I)	No (e.g. passive diffusion through membranes, solubility, lipophilicity, etc.)	Not relevant	Manual selection on the basis of researcher experience
2 (II)	Yes (e.g. ligand-receptor interactions)	Yes	Selection according to ligand's pose relatively to the target
3 (III)		No	Recommendation: select fragments from homogenous sets of compounds having a common scaffold and presumably, acting by the same mechanism

4 Fragment Selection and Assembling Strategies. Statistical Assessment of Calculated Contributions

Let's define *local* and *global* interpretation analysis. *Local* analysis is performed for one particular compound possessing several putative binding groups. *Local* analysis answers the question which of these groups is relatively more important and finds the most influencing fragment within the molecule. *Global* analysis consists of grouping fragment contributions of different molecules and comparing them to explain or extend experimentally observed trends in a structure-property relationship. *Local* analysis can be made for any compound of a data set, whereas the results of *global* interpretation depend on fragment selection and grouping strategy (Table 1). The simplest case (scenario 1 in Table 1)—no specific orientation of compounds relative to their target is expected or it can be disregarded (e.g., solubility, lipophilicity, passive diffusion through membranes, etc.). Fragment selection and grouping can be

guided by general considerations and depend on the decision of the researcher. If the orientation of compounds relative to their target is important and known from experimental studies or modeling (scenario 2 in Table 1), then this information should be taken into account during fragment selection and grouping. For example, if it is known that particular fragments of investigated compounds form H-bonds with the same amino acid residue, then these fragments can be grouped and analyzed together in order to obtain relevant interpretation results. In the worst case (scenario 3 in Table 1), the orientation of investigated compounds is important but unknown, we can analyze only homogenous sets of compounds tacitly assuming the identical interaction mode of all investigated compounds. MMP analysis (Leach et al. 2006) or SAR matrices (Wassermann et al. 2012) can also perform this analysis. Such data sets that comprise compounds with unknown but different mechanisms of action are the most common and therefore should be carefully analyzed.

A *global* analysis represents a contribution of each fragment by a distribution of its contributions for different molecules. To estimate fragment contributions' deviations from zero, the appropriate standard parametric/nonparametric statistical tests, such as the t-test, Wilcoxon rank test, etc. can be applied. However, this does not take into account a model error which may be quite large and may affect calculated contributions, e.g., if the model error is much greater than the difference between two predicted values, this difference can be a result of a chance correlation captured by the model. Therefore, to estimate practical significance we suggest not only to apply standard statistical tests but also to compare calculated contributions to model error estimated from cross-validation or from external testing. For example, we can express contribution values in units of the model error (units of RMSE for regression models) and choose a reasonable threshold value to separate significant and non-significant contributions.

5 Stereochemical Interpretation

As mentioned above, in $(2 + 0.X)$ SiRMS approach chiral 3D simplexes are used along with 2D simplexes. The use of 3D chiral simplexes can be considered significant if their addition to a 2D QSAR model improves its predictive performance. Number X is a sum of relative influences of chiral descriptors included in the final model. In the current study, relative influences of descriptors are defined by the corresponding regression coefficients in the PLS equation. The value of X can be interpreted as a factor describing the role of chiral features in a studied activity/property. For the molecules with multiple chiral centers, SiRMS approach allows to estimate relative influence of different chiral centers to the studied activity based on the following Formula (1):

$$RI_c = \sum_{i=1}^n \frac{Q_i}{4} \times Inf_i, \quad (1)$$

where RI_c is the relative influence of a particular chiral center, n is the number of chiral simplexes which include the selected chiral center, Q_i is the quantity of i -th simplex including this center, 4 is the number of atoms in each simplex and Inf_i is a relative influence of i -th simplex to 2.XD QSAR model calculated as modulo of regression coefficient of each descriptor in the model, %.

When it is necessary to calculate the overall relative influence of different elements of chirality (such as centers, axis or plane of chirality), one should add relative influences of chiral simplexes describing the corresponding type of chirality to the model. X in this case can be calculated as in (2)

$$X = Inf_c + Inf_a + Inf_p = \sum_{ic=1}^n Inf_{ic} + \sum_{ia=1}^n Inf_{ia} + \sum_{ip=1}^n Inf_{ip}, \quad (2)$$

where Inf_c , Inf_a and Inf_p are influences of central, axial and planar chirality, and Inf_{ic} , Inf_{ia} and Inf_{ip} are modulo of regression coefficients of i -th chiral simplex describing central, axial and planar chirality, respectively, included in the model. In the case if there is no particular type of chirality, the corresponding influence value is 0. This interpretation approach belongs to model-specific ones because it uses regression coefficients of a PLS model.

6 Structural and Physicochemical Interpretation. Case Studies

In this section we will: (i) compare results obtained by using the developed structural interpretation approach and classical Free-Wilson approach on the original data set used by Free and Wilson; (ii) demonstrate applicability of structural and physicochemical interpretation approaches on several real data sets which belong to different types according to the classification provided in Table 1. Since searching for the general trends in structure-property relationship comprises the greatest interest, we focus further analysis mainly on *global* interpretation. However, in some cases we demonstrate and discuss results of the *local* interpretation as well.

6.1 Comparison of the Approach of Structural Interpretation of QSAR Models with Classical Approaches

6.1.1 Hammett Constants Analysis

One of the first Linear Free-Energy Relationships appeared to be the Hammett equation. In the pioneering study Hammett, connected the acidities (dissociation constants) of substituted benzoic acids and the rates of alkaline hydrolysis of

substituted ethyl benzoates (Hammett 1937). From the received correlation they obtained the following equation:

$$\log(k/k_0) = \sigma\rho, \quad (3)$$

where σ —the substituent constant, ρ —reaction constant, k —the rate constant for the substituted compound, k_0 —reaction constant for the unsubstituted compound.

Unfortunately, in the original study, Hammett and coworkers presented no initial data on the dissociation constants for the listed 36 compounds, and in the further studies the dissociation constants are present only for the 17 substituents (taken from various sources). Thus, for the analysis we selected a study by Lindberg et al. (1976) and its further enhancement by Takahata and Chong (2005), where Core Electron-binding Energy (CEBE) shifts were chosen to obtain Hammett-like equation. The data set contained 29 different para-, ortho- and metasubstituted fluorobenzenes (Fig. 8), the values of CEBE shifts were expressed in eV. A direct comparison of fragment contributions was possible as the equation has a linear form.

$$\Delta CEBE = \kappa\sigma \quad (4)$$

$$\kappa = 2.3kT(\rho - \rho^*), \quad (5)$$

where κ —calculated parameter, σ —Hammett substituent constant, ρ and ρ^* —reaction constants specific for the neutral molecule and core ionized molecule, respectively.

Despite the small size of the data set, the model performance for the RF, SVM, GBM and PLS methods was quite high (Table 2). Calculated fragment contributions had an excellent agreement among different models and had a high correlation with the tabulated σ values ($R^2 = 0.85$ for the consensus model) (Takahata and Chong 2005). However the systematic shift between the calculated and tabulated

Fig. 8 Substituted fluorobenzenes used for modeling of Core electron-binding constants (Lindberg et al. 1976)

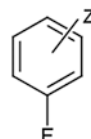


Table 2 Five-fold cross-validation statistical parameters of QSPR models for the critical volumes

Model	Q ²	RMSE, eV
GBM	0.81	0.20
RF	0.71	0.24
SVM	0.74	0.22
PLS	0.87	0.19
Consensus	0.88	0.18

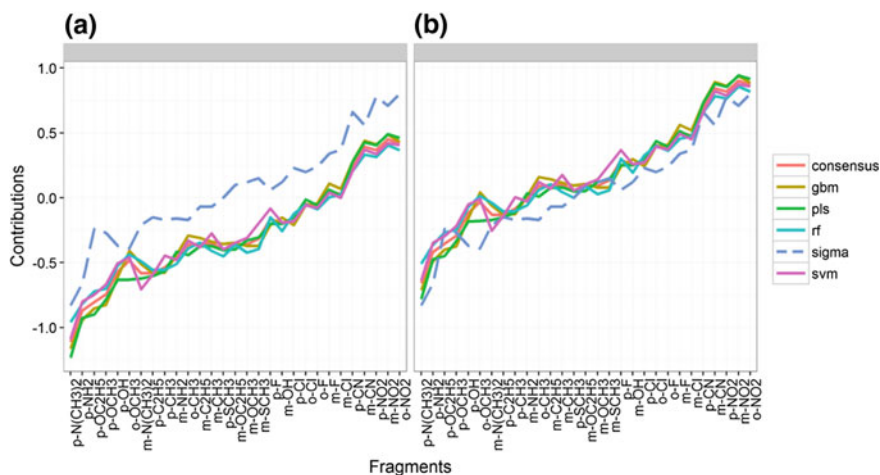


Fig. 9 The reference values of σ constants and **a** the original calculated fragment contributions and **b** the calculated fragment contributions with the regard for the contribution of the reference **methylsulfonyl** group

values was observed (Fig. 9a). It can be explained by the choice of a reference point. Lindberg et al. (1976) chose **1-fluoro-4-(methylsulfonyl)benzene** as the reference compound for the σ constants and assigned value of 0 to **methylsulfonyl group**. Thus, subtracting the calculated contribution value for the reference group results in a more coherent picture of the calculated contributions and tabulated values of σ constants (Fig. 9b).

Physicochemical interpretation of the consensus model reveals the effect of electrostatics term, which is the biggest effect for the strong electron donor substituents (Fig. 10). That means that these fragments have favorable distribution of charges and are more sensitive to their changes. Other substituents (particularly having moderate overall contributions) are more sensitive to the changes of electron polarizability and hydrophobicity. These conclusions partially correspond to the idea that Hammett constants represent effects of electron polarizability and partial charges distribution.

6.1.2 Free-Wilson Models

One of the data sets used by Free and Wilson for their pioneering work contained 29 compounds with associated LD_{50} values expressed in mg/10 g after i.p. injection in mice (Fig. 11) (Free and Wilson 1964). The weight units were not converted into molar units in order that the results would be comparable to published values of contributions.

The performance was poor for the five-fold cross-validation of RF, GBM, SVM and PLS models as could be expected due to the small size of the data set (Table 3).

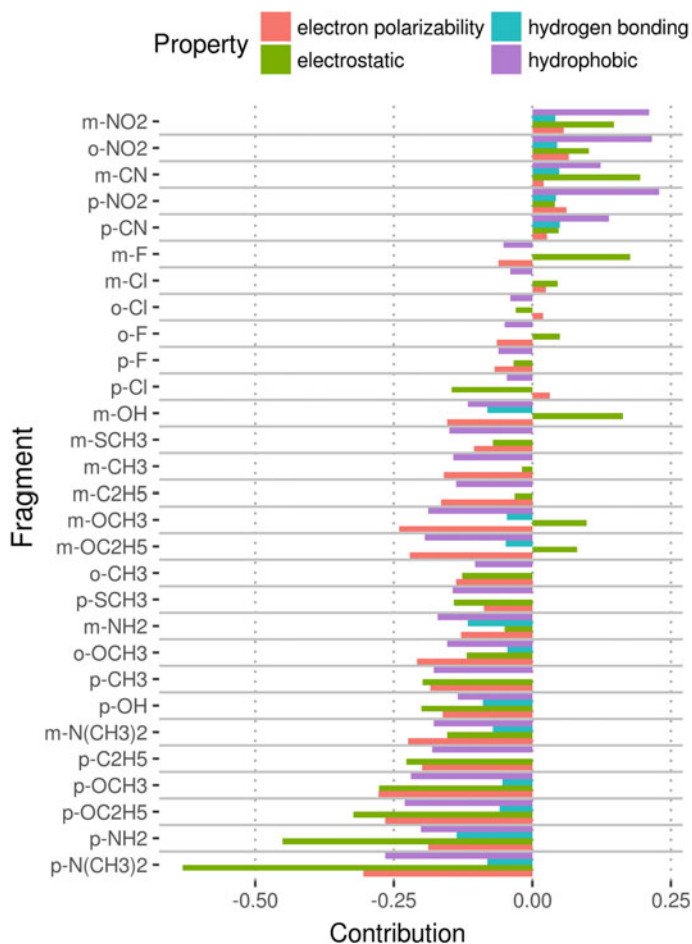
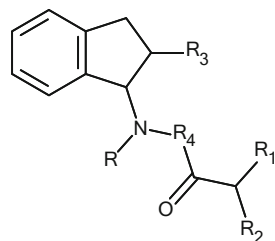


Fig. 10 Comparison of values for the sigma constants and the calculated substituent contributions (only descriptors weighted by the dispersive interactions)

Fig. 11 Structures of compounds of the data set from Free and Wilson paper (1964). R = H, CH₃; R₁ = H, CH₃, C₂H₅; R₂ = N(CH₃)₂, N(C₂H₅)₂, morpholino; R₃ = H, phenyl; R₄ = nothing, -CONH-



Despite the practical uselessness of the weak models, we performed an interpretation to compare results with the original Free-Wilson approach. As it can be seen

Table 3 Five-fold cross-validation statistical parameters of QSAR models for the Free-Wilson data set

	Q^2	RMSE
RF	0.34	1.47
GBM	0.33	1.48
SVM	0.43	1.37
PLS	0.26	1.56
Consensus	0.38	1.43

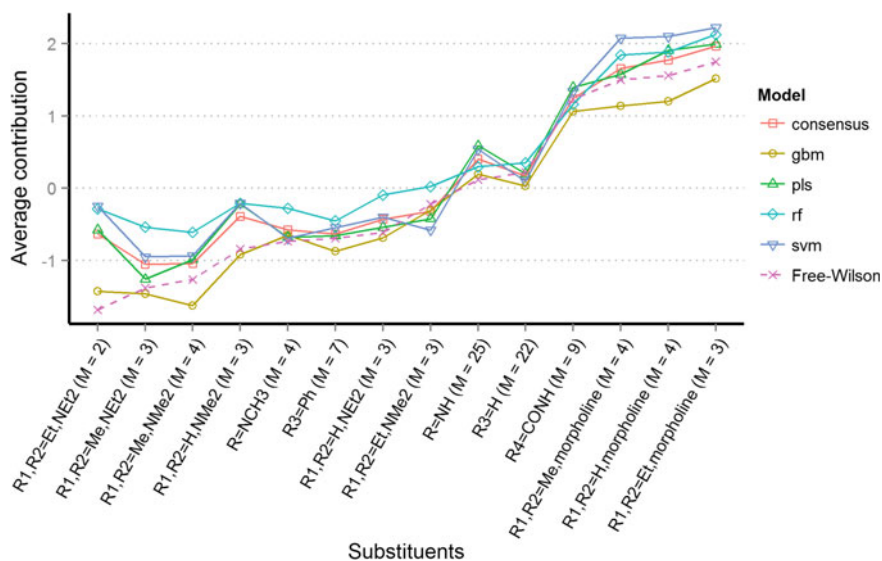


Fig. 12 Calculated average contributions of fragments for the Free-Wilson data set in comparison with results from original publication of Free and Wilson (Reprinted with permission from *J. Chem. Inf. Model.*, 2016, 56, 1455–1469. Copyright 2016 American Chemical Society.)

from Fig. 12, there is a good agreement between contributions calculated from different QSAR models and the contributions reported by Free and Wilson.

These two benchmark studies showed that the universal fragment interpretation approach possesses a consistency and inner logic, and the calculated fragment contributions proved comparable to those in previous classical studies.

6.2 Critical Properties of Organic Compounds (Type I Data Set)

Critical point—one of the most important physical characteristics of the compounds—describes the state on a phase surface where liquid and gas phases behave alike. Three parameters robustly and non-variantly characterize critical

point—critical temperature, critical pressure and critical volume. Traditional thermodynamic methods for the critical properties mainly base their predictions on the additive group-contributions (GC) schemes. In this case study we chose critical volume (V_c) as a property of interest and some of the most popular GC methods for comparison—method of Joback and Reid (1987), Marrero-Pardillo method (First-Order) and Marrero-Pardillo method (Second-Order) (Marrero-Morejón and Pardillo-Fontdevila 1999). Marrero-Pardillo method (Second-Order) suggests to use more complex fragments instead of using simple functional groups, i.e., secondary carbon in ring connected to a carbonyl. According to the authors, such approach allows to account for inter-group influences. The labels of fragments correspond to those in the original studies.

The experimental data were taken from the comprehensive handbook (Reid et al. 1987). Then wrong or incomplete data were cured using NIST Webbook database (Thermodynamics Research Center, NIST Boulder Laboratories, M. Frenkel director 2013). Among considered compounds were those of various classes, such as saturated and unsaturated hydrocarbons, aromatic hydrocarbons and their derivatives, heterocyclic compounds, alcohols, ethers, esters, various halogenated compounds, etc. The experimental V_c values were available for 309 compounds. The five-fold cross-validation showed very high performance for all models with Q^2 varying from 0.90 to 0.92. The RMSE values ranged from 38 to 43 cm^3/mol (Table 4).

Since there were no significant differences between calculated fragment contributions among different QSPR models, the results of only the consensus model will be discussed below. The main difficulty in comparing the interpretation results with reference group contribution values was comprised by the fact that every GC method was developed on a data set different from the studied one, and complete information about data sets is not available. We eliminated the fragments with insignificant contributions according to the Wilcoxon test ($\alpha = 0.95$). For the sake of visual comparison, the y-axis scale is the same for all plots. The analysis shows quite similar trends for Joback method with median fragment contributions calculated from the consensus model. The RMSE value between them is only 19 cm^3/mol (Fig. 13), and we connect the difference in contributions for a non-ring oxygen atoms with a discrepancy between the used data sets. The comparison with Marrero-Pardillo First-Order method (Fig. 13) shows less consistent results with QSPR method being more sensitive towards differences in the oxygen fragments (RMSE being 51 cm^3/mol). The comparison with Marrero-Pardillo Second-Order method (Fig. 13) demonstrates less divergence between results (RMSE is

Table 4 Five-fold cross-validation statistical parameters of QSPR models for the critical volumes

Model	Q^2	RMSE, cm^3/mol
GBM	0.90	43
RF	0.90	43
SVM	0.92	38
PLS	0.91	41
Consensus	0.93	35

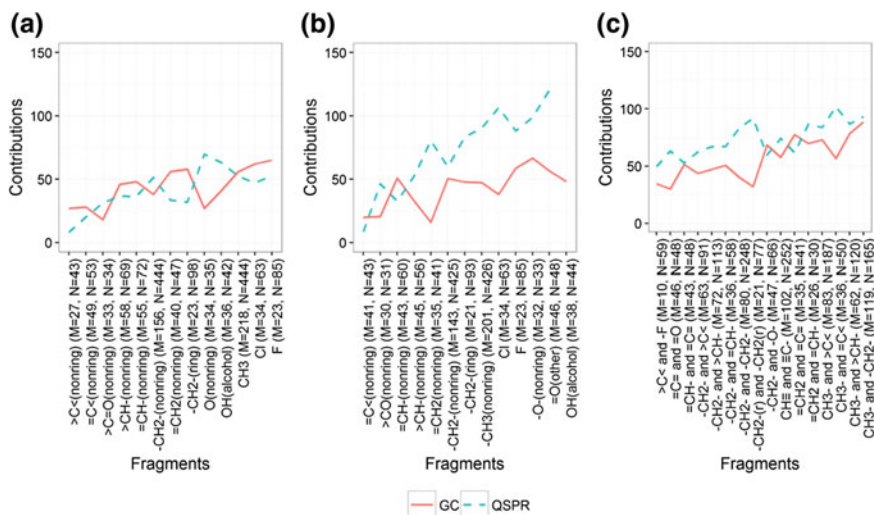


Fig. 13 Comparison of **a** Joback group contribution, **b** Marerro-Pardillo First-Order **c** and Marerro-Pardillo Second-Order values and the median values of the calculated fragment contributions from the consensus QSPR model. M is the number of compounds comprising a fragment. N is the number of fragments across the whole data set (some compounds have several identical fragments and their contributions were estimated separately). For Marerro-Pardillo Second-Order method (r) denotes ring group

26 cm³/mol), which can be explained by a more detailed description of the fragments and higher precision of the Second-Order method itself. These features consequently lead to statistically more reliable contributions and decrease influence of difference between the data sets.

In this case study, the structure-property relationships captured the trends of influence, previously discovered by the traditional group contribution methods. Calculated contributions values deviate from the group contributions, which we connect mainly with the difference in the data sets, as well as with the nature the QSPR models. Since the QSPR models possess higher mathematical complexity, they are also apt to describe complex relationships (unlike simpler group contributions methods). Predictive performance of the built QSPR models is very high and their applicability domain is rigorously defined, which makes use of QSPR models highly perspective for critical volumes prediction of organic compounds.

6.3 RGD-Mimetics—Antagonists of Fibrinogen Receptor (Type II Data Set)

Arg-Gly-Asp (RGD) sequence of fibrinogen responses for the interaction of fibrinogen with its receptor followed by a thrombus formation (Gartner and Bennett

1985; Andrieux et al. 1989). For a long time, researchers have focused on the development of antagonists of the fibrinogen receptor which mimic the RGD sequence (Hartman et al. 1992; Scarborough et al. 1993; Egbertson et al. 1994). The data set consisted of 325 antagonists of the fibrinogen receptor (RGD-mimetics) with a measured affinity value expressed as pIC_{50} values collected from ChEMBL database (Gaulton et al. 2012) and our own studies. The whole data set is published in the recent article (Polishchuk et al. 2016).

Each compound of this data set can be represented as consisting of three parts: Arg- and Asp-mimetics connected by a linker moiety (Fig. 14). We established ligand-protein interactions of these compounds in our previous docking studies (Polishchuk et al. 2015). Arg-mimetics containing a basic nitrogen atom form a charged H-bond with Asp224 and Ser225 residues of the fibrinogen receptor. Asp-mimetics contain a carboxylic acid residue which coordinates with Mg^{2+} ions inside the protein cavity and substituents which form H-bonds with Arg214 and Asn215. There are several hydrophobic residues in the binding site which may interact with Arg-mimetic and linker moiety of ligands. The linker part of ligands is

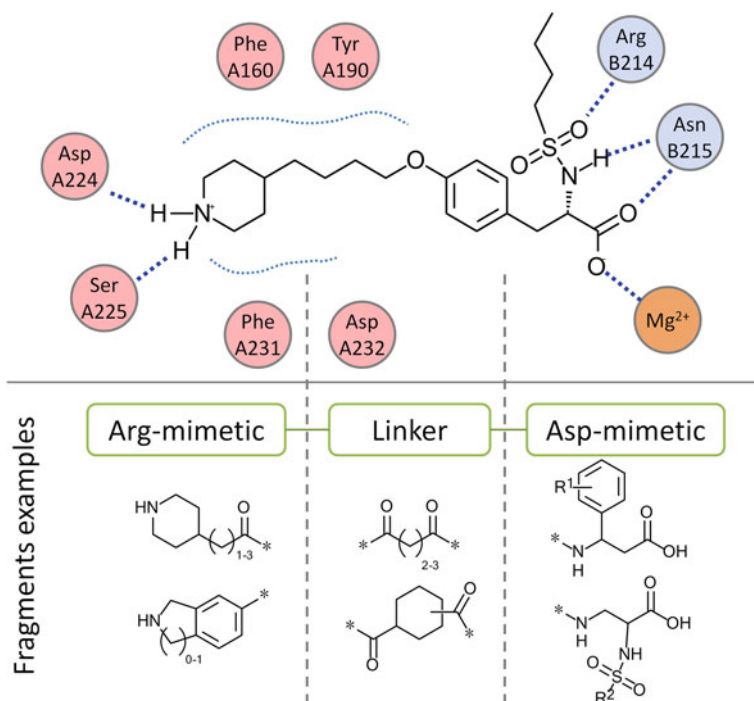


Fig. 14 Binding pattern of tirofiban—commercial antagonist of fibrinogen receptor (PDB code 2VDM) and the general representation of antagonists of the fibrinogen receptor which has Asp- and Arg-mimetic parts linked together, and several examples of corresponding fragments (Reprinted with permission from J. Chem. Inf. Model., 2016, 56, 1455–1469. Copyright 2016 American Chemical Society.)

Table 5 Five-fold cross-validation statistical parameters of QSAR models for the RGD-mimetics data set

	Q ²	RMSE
RF	0.72	0.81
GBM	0.68	0.86
SVM	0.70	0.82
PLS	0.67	0.88
Consensus	0.73	0.79

exposed and may be solvated or form H-bonds via water molecules with Asp232 residue (Polishchuk et al. 2015).

RF, GBM, SVM and PLS models were built for the compounds of this data set. They have satisfactory and comparable predictive performance (Table 5). As mentioned above, the compounds in this data set can be virtually split into three parts which interact with corresponding amino acid residues in the binding site of the fibrinogen receptor. Therefore, analysis of contributions was performed separately for these three groups of fragments (Fig. 15). The concordance between fragment contributions calculated across different models was high ($R = 0.89 - 0.98$). For this reason, analysis of interpretation results was done for the consensus model only (Fig. 16). Two-sided Wilcoxon rank test was applied to test the statistical significance of contributions. However, calculated contributions are affected by the accuracy and predictive performance of models. For this reason, it is feasible to compare contributions relative to a cross-validation error (RMSE). Contributions which are within 1 unit of RMSE can be considered insignificant and their analysis should be done with care.

Clear trends of structure-affinity relationship were observed for each group of fragments: Arg-mimetics, linkers and Asp-mimetics (Fig. 16). Cyclic secondary amines as Arg-mimetics increase affinity for the fibrinogen receptor more than pyridyl, amidino or guanidino groups. Unexpectedly, there were five outliers in the L3 linker group. More thoughtful analysis revealed that all five cases correspond to the most active compounds in the data set. It points to some restrictions due to QSAR modeling. Models which cannot extrapolate (e.g., such as RF or GBM) always return predicted values within the range of observed values of the training set compounds. This causes contributions of almost any fragment comprising the most active compounds to be positive. The same is true for the contribution of fragments in the least active compounds. Thus, special attention should be paid to such compounds and their analysis.

Asp-mimetics were the most diverse part of RGD-peptidomimetics. The frequently occurring D8 fragment has a very large range of contributions values due to the substantial influence of molecular context. At the same time, fragment D6 (also present in different molecular contexts) has a smaller range of contributions. In general, the variance of contribution values of Arg-mimetics is substantially smaller than that for linkers and Asp-mimetics. This indicates that the nature of Arg-mimetics may be more important for binding to fibrinogen receptor than linkers and Asp-mimetics whose contributions are highly context dependent. This

Arg-mimetic						
R1	R2	R3	R4	R5	R6	R7
Linker						
L1	L2	L3	L4	L5	L6	L7
Asp-mimetic						
D1	D2	D3	D4	D5	D6	D7
D8	D9	R=OCH ₃ , OCH ₂ O				

Fig. 15 Most frequently occurring fragments in compounds of the RGD-peptidomimetics data set (Reprinted with permission from J. Chem. Inf. Model., 2016, 56, 1455–1469. Copyright 2016 American Chemical Society.)

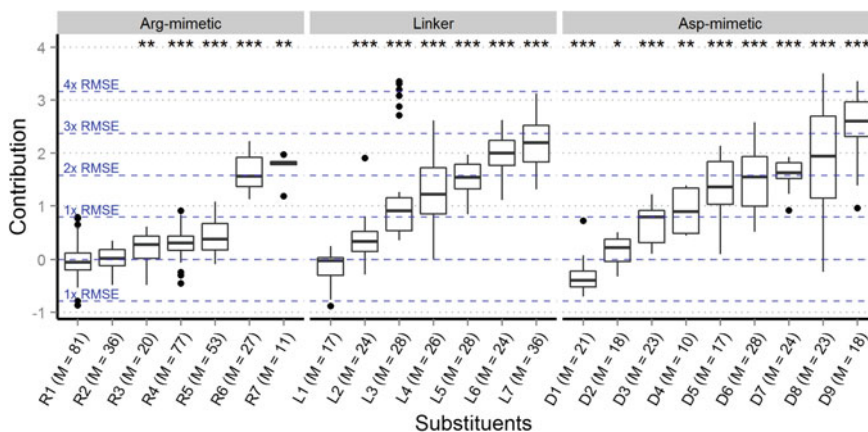


Fig. 16 Distribution of overall fragments contributions of RGD-mimetics calculated from the consensus QSAR model. M is the number of compounds comprising a fragment. Asterisks refer to statistical significance calculated by the two-sided Wilcoxon rank test (p-value): *** <0.001, ** <0.01, * <0.05 (Reprinted with permission from J. Chem. Inf. Model., 2016, 56, 1455–1469. Copyright 2016 American Chemical Society.)

information along with established trends in structure-property relationship may be used for drug design per se or as a guideline for the researcher.

It is important to note that interpretation results depend on available data sets and this should always be taken into account in any analysis. For example, it is well known from pharmacophore modeling, docking and X-ray data of ligand-protein complexes that fibrinogen receptor antagonists should contain positively and negatively charged groups at a distance of 15–20 Å (Polishchuk et al. 2015; Hartman et al. 1992; Egbertson et al. 1994; Springer 2008). These groups are essential for binding to the receptor. However, according to interpretation of results, contributions of some Arg-groups are close to zero and not statistically significant (Fig. 16). Thus, we can erroneously conclude that these groups are not very important for binding. But such results are easily explained by the biased data set which does not contain true non-binders. Compounds with guanidine (R1) and pyrimidine (R2) groups have the lowest affinity values (10–100 μM) in the data set and therefore their contributions are very low. However, that does not mean that these groups are unimportant for receptor recognition.

The *global* and *local* physicochemical interpretation reveals the high contribution of the electrostatic term (Fig. 17), which we assume the main driving force of ligand-receptor interaction. This assumption can be supported by the following considerations: (1) ligands have at least one positively and one negatively charged group which is essential for ligand-receptor recognition (Fig. 18); (2) there are commonly one or two charged H-bonds in ligand-receptor complexes according to earlier molecular docking studies (Fig. 18), (3) the desolvation effect of ligands

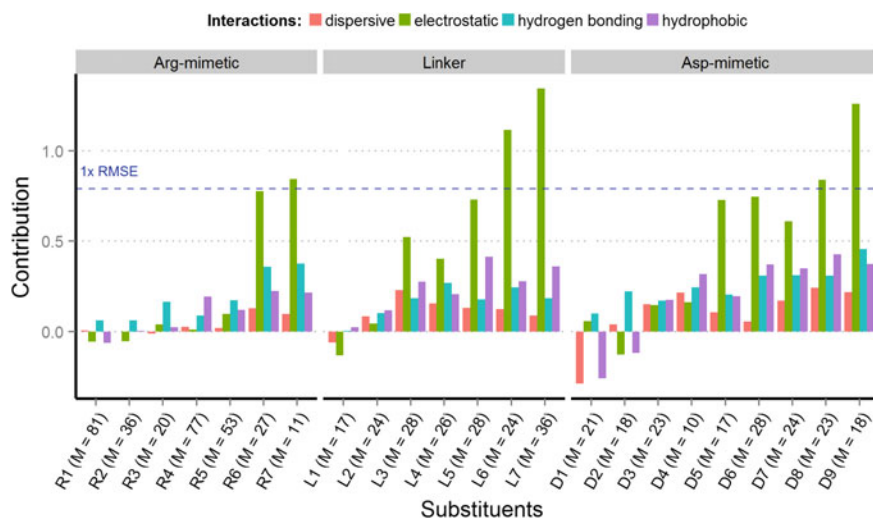


Fig. 17 Median contributions of physicochemical terms in ligand-protein interactions of the RGD-mimetics data set compounds calculated from the consensus model consisting of RF, GBM, SVM and PLS models. Definition of M was given in the caption of Fig. 13 (Reprinted with permission from J. Chem. Inf. Model., 2016, 56, 1455–1469. Copyright 2016 American Chemical Society.)

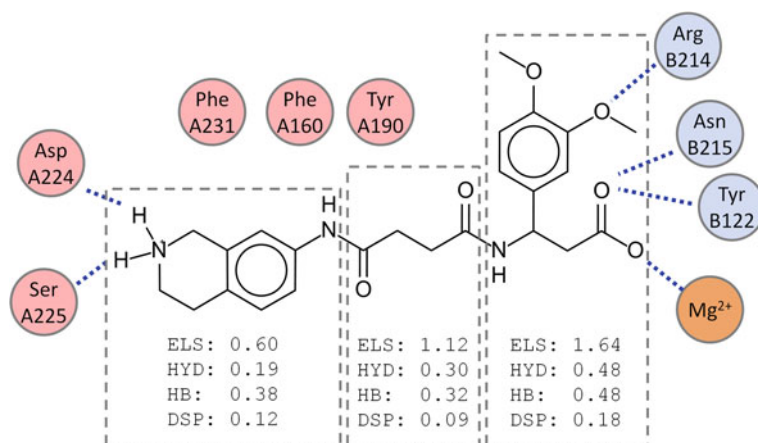


Fig. 18 Interaction map of a selected ligand with the fibrinogen receptor and calculated contributions of physicochemical terms for separate fragments from the consensus model (ELS—electrostatic, HYD—hydrophobic, HB—hydrogen bonding, DSP—dispersive). This is an example of *local* interpretation (Reprinted with permission from J. Chem. Inf. Model., 2016, 56, 1455–1469. Copyright 2016 American Chemical Society.)

which depends on the distribution of partial atomic charges can also play an important role in ligand binding and cannot be estimated directly. Less significant effects of hydrogen bonding (relative to the electrostatic term) may be explained by the charged nature of formed H-bonds between ligands and Asp224 and Arg214 residues of the fibrinogen receptor. There are few hydrophobic residues in the binding pocket and correspondingly, relatively small contributions of hydrophobic effects of fragments are observed in the consensus QSAR model. The contributions of dispersive interactions are the smallest as these forces are usually very weak and do not substantially influence affinity values.

6.4 Acute Oral Toxicity in Rats (Type III Data Set)

The data were obtained from the T.E.S.T. 4.1 program (Toxicity Estimation Software Tool) provided by U.S. EPA (Environmental Protection Agency) (<http://www2.epa.gov/chemical-research/toxicity-estimation-software-tool-test>). LD₅₀ values were converted from mass to molar units and expressed as $-\log\text{LD}_{50}$ (LD₅₀, mol/kg). 7205 compounds remained in the data set after removal of salts, undefined isomeric mixtures, polymers and mixtures.

Statistical characteristics of individual RF, GBM, SVM and PLS models and their consensus predictions are given in Table 6. Despite the poor predictive performance of the PLS model, we included it in the consensus model, and the predictive ability of the latter, and the interpretation results remained unchanged. Since the consensus model provides averaged and thus less biased predictions we

Table 6 Five-fold cross-validation statistical parameters of QSAR models for the acute oral toxicity data set

	Q ²	RMSE
RF	0.61	0.59
GBM	0.56	0.63
SVM	0.54	0.64
PLS	0.44	0.71
Consensus	0.60	0.60

performed only its interpretation. However, the calculated contributions among individual models were also in a good agreement.

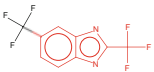
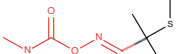
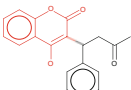
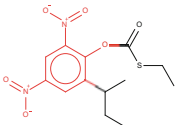
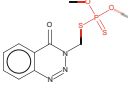
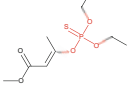
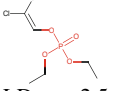
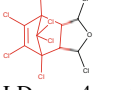
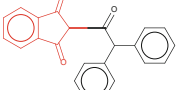
First, the contributions of known toxicophores with established mechanism of action were calculated to confirm the ability of the interpretation approach to rank them correctly relatively to other structural motifs (Table 7). The mentioned toxicophores ranked top among all considered fragments (Fig. 19). We analyzed an influence of a molecular context and distribution of values for some toxicophores with common structures such as the carbamate group. The contributions of O-(methylaminocarbonyl)oxime exceeded the contribution of the parent carbamate group. But the cyclic carbamate fragment proved virtually non-toxic (Table 7).

Second, we analyzed the contributions of other highly ranked fragments from the list of common functional groups and ring systems in order to find new potential toxicophores. Some, like nitrosamine and aziridine are known mutagens as it was shown by Kazius with co-workers (Kazius et al. 2005) and in our recent publication (Polishchuk et al. 2013). Others like piperazine and piperidine are frequently used in medicinal chemistry and not well known as toxicophores. The analysis of the molecular context of these groups revealed 4-phenylpiperazine and 4-phenylpiperidine moieties as probable toxicophores as they have significantly higher contribution values relative to the contributions of corresponding parent fragments. Halogens per se have relatively small contribution to acute toxicity and may be considered weak toxicophores. Fragments with negative contributions like carboxylic or sulfonic acid groups can be considered detoxicophores.

6.5 Comparison of BBB Permeability with Passive Diffusion Measured by PAMPA and P-gp Binding

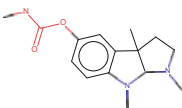
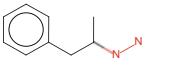
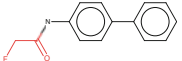
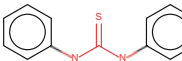
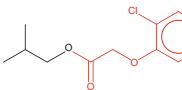
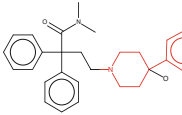
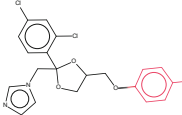
The transport of molecules across the blood brain barrier (BBB) is a highly restricted and controlled process. The tight junctions between adjacent cells, lack of capillary fenestration and low pinocytotic activity hamper transport across BBB. The major pathway for compounds to cross the BBB is the transcellular route, which depends on their physical-chemical characteristics and interactions with transporter proteins. Passive diffusion is the primary process of translocation from the blood stream to the brain for most drugs. For a given membrane or barrier

Table 7 List of fragments with known or presumed mechanism of toxic action* (Reprinted with permission from J. Chem. Inf. Model., 2016, 56, 1455–1469. Copyright 2016 American Chemical Society)

Fragment	Representative structures	Mode of action	References
<i>Known toxicophores</i>			
2-trifluoromethyl benzimidazole	 LD ₅₀ = 9.01 mg/kg	Inhibitors of oxidative phosphorylation	(Beechey 1966)
O-(methylamino-carbonyl)oxime	 LD ₅₀ = 0.5 mg/kg	Acetylcholine esterase inhibitors	(Čolović et al. 2013)
4-hydroxycoumarin	 LD ₅₀ = 1.6 mg/kg	The anticoagulant (vitamin K antagonists)	(Littin et al. 2000)
Dinitrophenoxy	 LD ₅₀ = 147 mg/kg	Inhibitors of oxidative phosphorylation	(Terada 1990; Grundlingh et al. 2011)
Phosphorodithioate	 LD ₅₀ = 7 mg/kg	Acetylcholine esterase inhibitors	(Čolović et al. 2013)
Phosphorothionate	 LD ₅₀ = 0.14 mg/kg		
Phosphoryl	 LD ₅₀ = 2.5 mg/kg		
Hexachlorononbornene	 LD ₅₀ = 4 mg/kg	GABA-gated chloride channel antagonists	(Casarett and Klaassen, 2008)
1,3-indandione	 LD ₅₀ = 1.5 mg/kg	The anticoagulant (vitamin K antagonists)	(Valchev et al. 2008)

(continued)

Table 7 (continued)

Fragment	Representative structures	Mode of action	References
Carbamate	 LD ₅₀ = 4.5 mg/kg	Acetylcholine esterase inhibitors	(Čolović et al. 2013)
NHNH ₂ (not hydrazide)	 LD ₅₀ = 34 mg/kg	Inhibition of GABA synthesis	(Medina 1963; O'Brien et al. 1964)
2-fluoroacetyl	 LD ₅₀ = 10 mg/kg	Inhibition of Krebs cycle	(Proudfoot et al. 2006)
Thiourea**	 LD ₅₀ = 50 mg/kg	Inhibitor of thyroid peroxidase	(Davidson et al. 1979)
2-(2,4-dichloro-phenoxy)acetyl	 LD ₅₀ = 300 mg/kg	Generation of free radicals, increases the lipid peroxidation process, depletion of ATP	(Bukowska 2006)
<i>Potential new toxicophores</i>			
Phenylpiperidine	 D ₅₀ = 98 mg/kg		
Phenylpiperazine	 LD ₅₀ = 166 mg/kg		

*SMARTS patterns of all analyzed fragments are given in Table S1 in Supporting materials (Polishchuk et al. 2016)

**The acute toxicity of thiourea varies with species, strain, age and iodine content of the diet. [Thiourea. Concise International Chemical Assessment Document, 49. World Health Organization, Geneva, 2003]

characteristics (surface area, water pores, tight junction diameters, histomorphology), the physicochemical parameters of molecules represent the major rate determinant for passive diffusion. Hitherto it is the most computable and measurable property which can be used as a predictor of BBB permeation. It depends on simple parameters that make sense for medicinal chemists (molecular weight, H-bond capacity, rotatable bonds, solvent-accessible surface area). Polar molecules are generally poor CNS agents unless they undergo active transport to pass the CNS. Substrates of P-gp are generally chemically unrelated drugs. The efflux transport depends on the drug binding to P-gp, the level of expression of P-gp, the

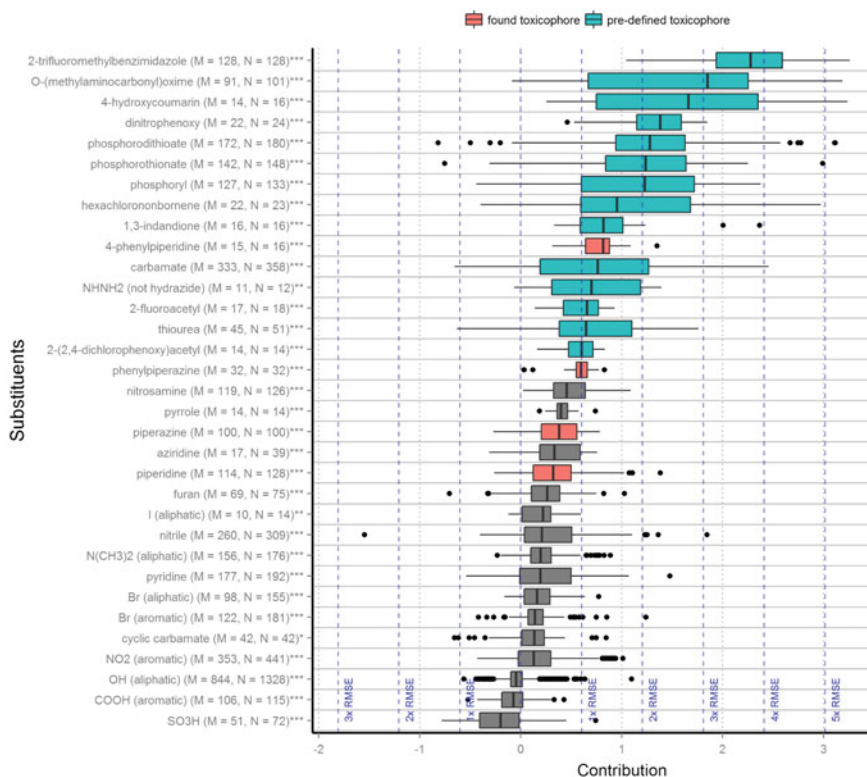


Fig. 19 Contributions to toxicity of different molecular fragments calculated on the basis of the consensus QSAR model. Definitions of M and N were given the caption of Fig. 13 (Reprinted with permission from *J. Chem. Inf. Model.*, 2016, 56, 1455–1469. Copyright 2016 American Chemical Society.)

drug concentration, and a constant equilibrium. The presence of P-gp inhibitors or modulators could modify this transport component significantly. Practically, for a majority of drugs the resulting BBB permeation can be represented as a sum of passive diffusion and efflux transport components (Adenot and Lahana 2004; Ma et al. 2005). Moreover, the inclusion of P-glycoprotein transport as a component of BBB permeation allows a clear mapping of drugs in terms of their behavior toward blood-brain barrier permeation. To estimate passive diffusion, the parallel artificial membrane permeability assay (PAMPA) test is frequently used (Kansy et al. 1998). Therefore, it would be reasonable to compare results of QSAR model interpretation for these three end-points: permeability through BBB and PAMPA and substrate binding to P-gp.

For this study three different data sets were collected:

- (i) 321 compounds with measured BBB permeability (178 permeable and 143 non-permeable); these compounds permeate by mainly passive diffusion (Polishchuk et al. 2016)

- (ii) 281 drug substances with measured passive permeability by Double Sink PAMPA (141 permeable and 140 non-permeable) (Avdeef 2012)
- (iii) 194 compounds with associated data on P-gp (98-substrate and 96-non-substrate) (Bikadi et al. 2011)

All data sets had common compounds: BBB/PAMPA 54, BBB/P-gp 45, P-gp/PAMPA 82 and 30 compounds were common for all data sets.

Table 8 Five-fold cross-validation performance for BBB, PAMPA and P-gp models

	Balanced accuracy	Sensitivity	Specificity	Kappa
<i>BBB permeability</i>				
RF	0.76	0.81	0.71	0.52
GBM	0.77	0.84	0.69	0.54
SVM	0.75	0.79	0.70	0.49
Consensus	0.76	0.83	0.69	0.52
<i>P-gp binding</i>				
RF	0.75	0.76	0.75	0.51
GBM	0.76	0.74	0.78	0.53
SVM	0.70	0.62	0.78	0.40
Consensus	0.76	0.73	0.79	0.52
<i>PAMPA permeability</i>				
RF	0.80	0.79	0.81	0.60
GBM	0.82	0.82	0.81	0.64
SVM	0.80	0.77	0.83	0.59
Consensus	0.81	0.81	0.82	0.63

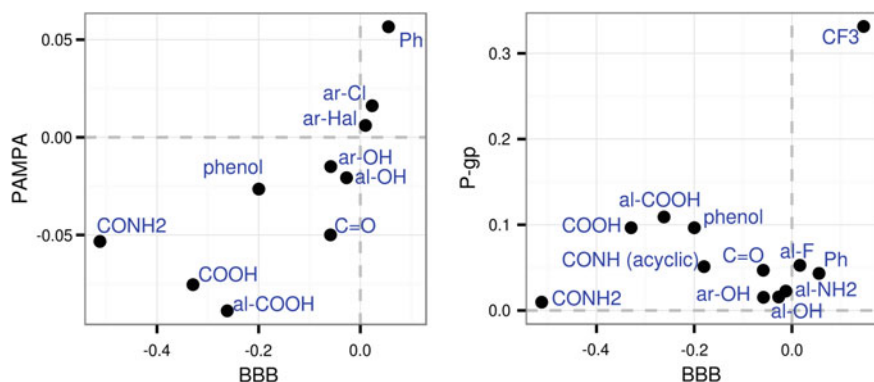


Fig. 20 Median contributions of fragments estimated from BBB, PAMPA and P-gp consensus models which occurred at least 7 times in both data sets. In the labels “al” means aliphatic atom, “ar” means aromatic atom

2D QSAR models were built by means of GBM, SVM and RF methods as well as the consensus models which included all corresponding individual models. All models had reasonable performance according to the five-fold cross-validation strategy (Table 8).

The structural interpretation results for all corresponding individual and consensus models were in good agreement, further we discuss the interpretation results of the consensus models only. One could expect that the similar interpretation results for BBB and PAMPA end-points because the BBB data set includes mainly compounds permeating by passive diffusion. Indeed, the plot of median contribution of the common fragments shows such trend (Fig. 20). The fragments which enhance PAMPA permeability also enhance BBB permeability and vice versa. Analysis of fragment contributions relationship to BBB permeability and P-gp binding proved more difficult. No clear trend exists between fragment contributions (Fig. 20). Some groups, like carboxylic, increase the chance of binding to P-gp and at the same time they have poor passive permeability. Therefore, compounds bearing such groups may have decreased BBB permeability due to both factors.

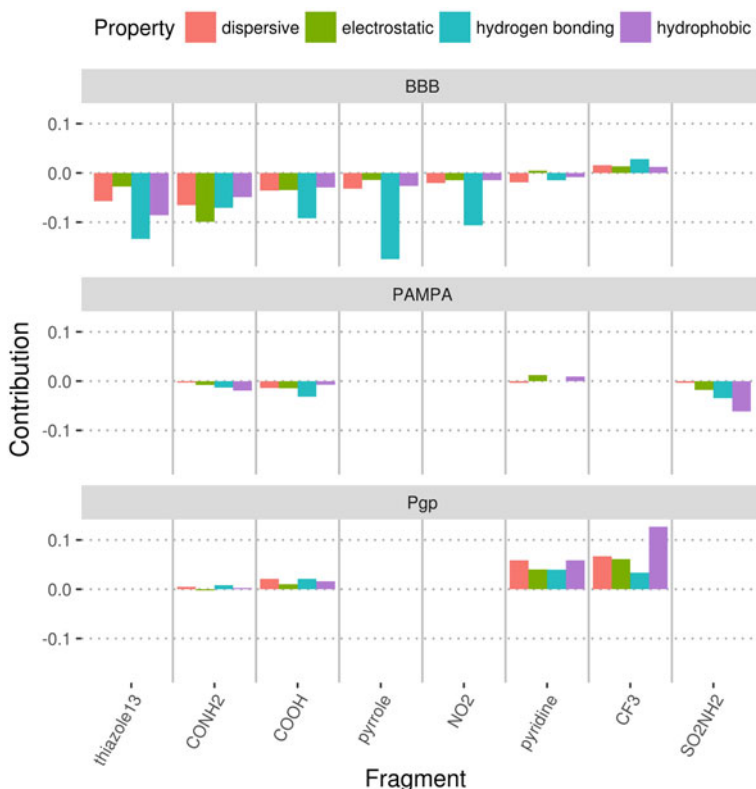


Fig. 21 Median physicochemical contributions of the most influential fragments from BBB, PAMPA and P-gp data sets

Other groups, like CF_3 , may increase passive diffusion and at the same time increase the chance of efflux. Therefore, researchers should find the trade-off between these two effects for compounds having such groups.

For many fragments physicochemical interpretation of the consensus models returns very small values, therefore we show the contributions of the most influential fragments only (Fig. 21). Negative influence of fragments with heteroatoms on BBB permeability can be explained by unfavorable hydrogen bonding. The positive impact of CF_3 group on P-gp binding presumably relates to the favorable hydrophobic interactions, whereas unfavorable hydrophobic interaction of SO_2NH_2 group results in its negative influence on passive diffusion measured by PAMPA. Generally these conclusions correspond to the experimentally observed relationships. Low BBB permeability may be caused by a large number of H-bond donors/acceptors and a large polar surface area (Hitchcock and Pennington 2006; Wager et al. 2010a, b; Ghose et al. 2012; Hitchcock 2012).

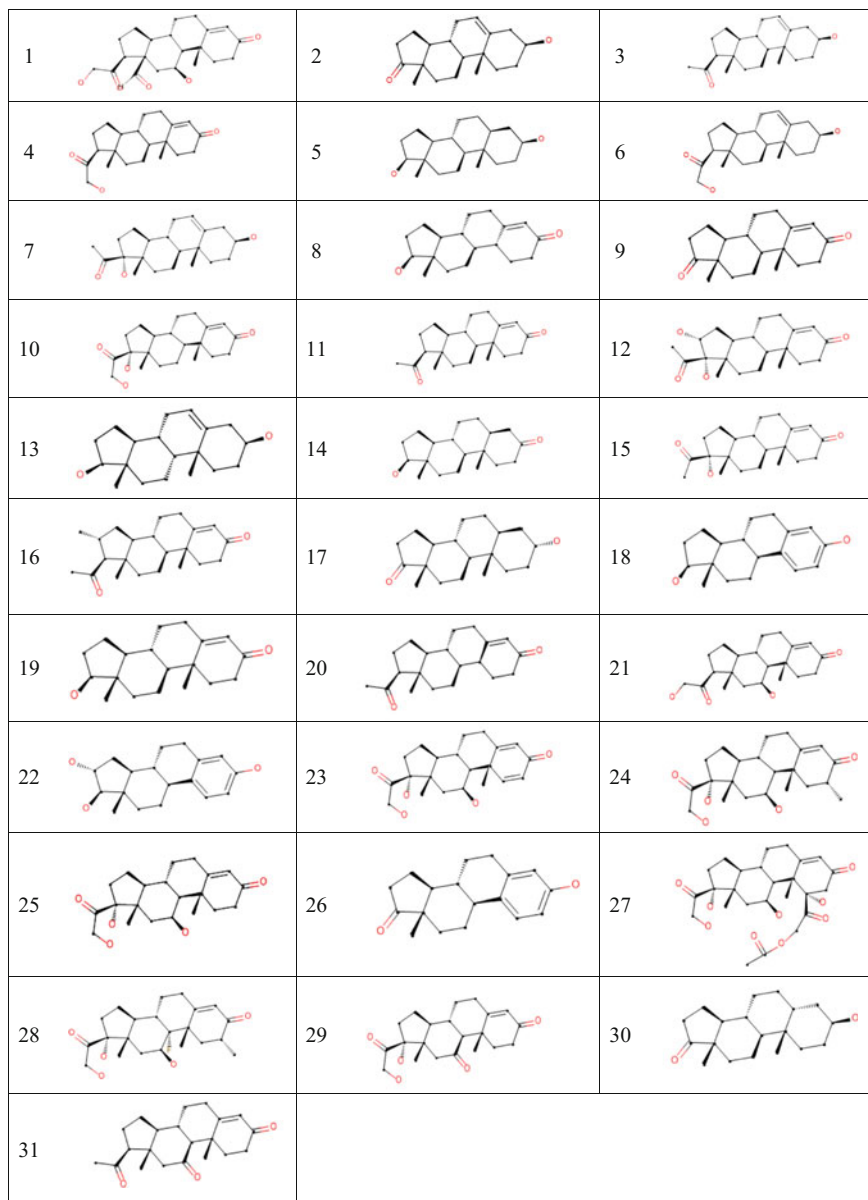
7 Stereochemical Interpretation. Case Studies

Two QSAR tasks were selected to demonstrate the applicability of (2 + 0.X)D SiRMS approach to stereochemical interpretation. The first data set contained compounds with multiple chiral centers and the second set contained compounds with chiral centers and axis of chirality. The (2 + 0.X)D SiRMS approach makes it possible to consider stereochemical features of those compounds and provide stereochemical interpretation. Both of the data sets belong to the type III data set, in which compound binding poses are important but unknown. But the data sets contain only congeneric series of compounds that makes interpretation reasonable assuming that they have the same binding mode.

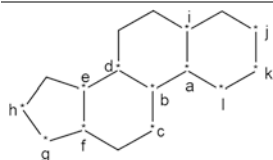
7.1 Cramer Steroids with Multiple Chiral Centers

The first data set—also known as the Cramer set—contained steroids with associated affinity values for corticosteroid-binding globulin (CBG) expressed as pK_a , where K_a is the association constant between steroid and CBG (Dunn et al. 1981; Cramer et al. 1988) (Fig. 22, Table 9). Many 3D-QSAR methods (Cramer et al. 1988; Silverman and Platt 1996; Lobato et al. 1997; Parretti et al. 1997; Besalú et al. 2002; Liu et al. 2002; Marrero-Ponce et al. 2008) use this data set as a benchmark. The studied compounds contain multiple chiral centers.

In Table 9 the numbers of chiral centers were assigned according to their positions in the scaffold (a-1) taking into account their molecular context. For example an atom in the same position as center 1 can become center 13, 15 or 18 according to the changes in the nearest surrounding. To describe chirality of each center, we calculated the set of connected 3D-simplexes including chiral center and

**Fig. 22** Structures of Kramer steroids

4 nearest atoms. Along with 2D-simplexes, these simplexes represent the structure of whole molecule and were used to build a PLS model. The obtained (2 + 0.X) D-SiRMS model has shown high quality, its statistical parameters exceed most of the known for 3D-QSAR models. In all previous studies only R^2 and Q^2

Table 9 The scaffold of the studied steroids with labeled stereogenic centers and affinity of compounds for corticosteroid-binding globulin

Scaffold of studied steroids and their stereogenic centers

№	pKa	Numbers and stereochemical configurations (R/S) of stereogenic centers											
		a	b	c	d	e	f	g	h	i	j	k	L
1	-6.27	1R	2S	3S	4S	5S	6R	7S	-	-	-	-	-
2	-5	1R	2S	-	4R	5S	8S	-	-	-	9S	-	-
3	-5.25	1R	2S	-	4S	5S	8S	7S	-	-	9S	-	-
4	-7.2	1R	2S	-	4R	5S	8S	7S	-	-	-	-	-
5	-5	1R	2S	-	4R	5S	8S	10S	-	11S	9S	-	-
6	-7.65	1R	2S	-	4S	5S	8S	7S	-	-	9S	-	-
7	-5	1R	2S	-	4R	5S	8S	12R	-	-	9S	-	-
8	-6.14	13R	2S	-	4R	5S	8S	10S	-	-	-	-	-
9	-5.76	1R	2S	-	4R	5S	8S	-	-	-	-	-	-
10	-7.88	1R	2S	-	4S	5S	8S	12R	-	-	-	-	-
11	-7.38	1R	2S	-	4S	5S	8S	7S	-	-	-	-	-
12	-6.24	1R	2S	-	4R	5S	8S	12S	14R	-	-	-	-
13	-5	1R	2R	-	4R	5S	8S	7S	-	-	9S	-	-
14	-5.91	15S	2S	-	4R	5S	8S	10S	-	11S	-	-	-
15	-7.74	1R	2S	-	4R	5S	8S	12R	-	-	-	-	-
16	-7.12	1R	2S	-	4S	5S	8S	7S	16R	-	-	-	-
17	-5.61	15S	2S	-	4R	5S	8S	-	-	11S	9R	-	-
18	-5	-	17S	-	4R	5S	8S	10S	-	-	-	-	-
19	-6.72	1R	2S	-	4R	5S	8S	10S	-	-	-	-	-
20	-6.81	13S	2S	-	4R	5S	8S	7S	-	-	-	-	-
21	-7.88	1R	2S	3S	4S	5S	8S	7S	-	-	-	-	-
22	-5	-	2S	-	4R	5S	8S	10S	14R	-	-	-	-
23	-7.51	18R	2S	3S	4S	5S	8S	12R	-	-	-	-	-
24	-7.68	1R	2S	3S	4S	5S	8S	12R	-	-	-	19R	-
25	-7.88	1R	2S	3S	4S	5S	8S	12R	-	-	-	-	-
26	-5	-	2S	-	4R	5S	8S	-	-	-	-	-	-
27	-7.55	1R	2S	3S	4S	5S	8S	12R	-	-	-	-	20R
28	-5.79	1S	21R	3S	4S	5S	8S	12R	-	-	-	19R	-
29	-6.89	1R	2S	-	4S	5S	8S	12R	-	-	-	-	-
30	-5.25	15S	2S	-	4R	5S	8S	-	-	11R	9S	-	-
31	-6.77	1R	2S	-	4S	5S	8S	7S	-	-	-	-	-

Table 10 Comparison of the statistical parameters for QSAR models built on Kramer steroids set

N _g	Descriptors	Level of structure representation	Statistical method	R ²	Q _{LOO} ²
1	(2 + 0.X)D-SiRMS	2.XD	PLS	0.90	0.84
2	2D-SiRMS	2D	PLS	0.86	0.80
3	3D-chiral TOMOCOMD-CARDD (Marrero-Ponce et al. 2008)	2.5D	MLR	0.95	0.83
4	MEDV (Liu et al. 2002)	3D	MLR + GA	0.86	0.77
5	TQSI (Lobato et al. 1997)	3D	MLR	0.83	0.76
6	CoMSIA (Parretti et al. 1997)	3D	PLS	0.76	0.73

Table 11 Relative influence of different chiral centers on CBG affinity of Kramer steroids

Chiral center	Configuration	Number of appearances	Average influence, %
1	R	21	2.81
2	S	28	2.51
3	S	7	1.41
4	R	17	1.54
4	S	13	1.40
5	S	31	2.46
7	S	10	3.56
8	S	30	3.68
9	S	7	1.22
10	S	6	2.90
11	S	3	2.48
12	R	9	2.85
15	S	3	2.73

(determination coefficient of leave-one-out cross validation) values were provided (Table 10).

The obtained (2 + 0.X)D-SiRMS model exceeds the ordinary 2D-SiRMS model in terms of predictive ability estimated additionally using five-fold cross-validation strategy: $Q_{5CV}^2 = 0.90$ and $RMSE_{5CV} = 0.34$ versus $Q_{5CV}^2 = 0.86$ and $RMSE_{5CV} = 0.39$, correspondingly. Thus, CBG affinity of those steroids is chirality-dependent and the use of stereochemical descriptors improves model performance. Stereochemical interpretation showed that relative influence of chiral descriptors (X in '2 + 0.X' D formula) is 29%. Thus, this task was solved by the 2.29 D-QSAR model.

During the computations we omitted those chiral centers that appeared less than 3 times in the data set (Table 11). There are centers with notably higher influence compared to others such as centers 7, 10 and 12 defining orientation of substituent in position g (Table 9) and centers 1, 15 and 8 which define configuration of centers

a and f, respectively. These centers define the coupling of cycles in steroid scaffolds. Analysis of overall influence of chiral centers grouped by position in the scaffold (a-l) shows that they can be ranked by an influence in the following order: $f > g > a > b > e$. All of the chiral centers have non-zero influence on the affinity according to this model. It proves that stereoconfigurations of all centers contribute to the binding of steroids to CBG.

7.2 *Data Set of Compounds with Mixed Central and Axial Chirality*

This data set included 40 naphthylisoquinoline alkaloids with antiplasmodial activity determined using the K1 strain (resistant to chloroquine and pyrimethamine), for which a modification of the [^3H]-hypoxanthine incorporation assay (Bringmann and Rummey 2003). Activity values were expressed as $\log\text{IC}_{50}$ where IC_{50} was measured in nmol/mL (Table 12). These compounds along with others were studied by 3D QSAR models via CoMSIA approach (Bringmann and Rummey 2003).

These compounds possess two types of chirality—central and axial. The feature vector representing those compounds thus should contain (i) ordinary 2D simplex descriptors, (ii) 3D simplex descriptors representing chiral centers as shown in the previous study and (iii) 3D simplex descriptors representing axial chirality. We employed a system of unconnected simplexes containing atoms pairs located in different planes from the axis of chirality to describe axial chirality and the connected simplexes to describe atoms at chiral axis and next to them (Fig. 23). All of the atom pairs in unconnected simplexes are not planar and can be used to generate chiral descriptors. All of the connected simplexes include C-C bond which is equal to the axis of chirality.

The built PLS models were validated according to the five-fold cross-validation protocol. The $(2 + 0.X)\text{D}$ SiRMS model had satisfying statistical parameters ($R^2 = 0.81$, $Q^2 = 0.74$, $R_{5\text{CV}}^2 = 0.78$, $S_{5\text{CV}} = 0.33$) and exceeded those for the 2D SiRMS model ($R^2 = 0.80$, $Q^2 = 0.63$, $R_{5\text{CV}}^2 = 0.72$, $S_{5\text{CV}} = 0.38$). Thus, this 2.X-D QSAR model can be used for stereochemical interpretation. It proves impossible to compare our results with the QSAR studies from Bringmann and Rummey (2003) because the authors included in their training set compounds with stereochemically unstable axis—axis around which free rotation is possible. However, predictive ability for biaryl compounds in Bringmann and Rummey (2003) was not quite high compared to this study— $R^2(\text{ts}) = 0.59$.

The relative influence of simplex descriptors for axial chirality is 12% (including 8.5% for unconnected simplexes and 3.5% for connected ones) and 5% for the simplex descriptors represented central chirality. Thus, one can assume that this model is at 2.17D level. Since descriptors for both types of chirality contribute to the model this evidences that all types of stereochemistry are important for anti-malarial activity. Larger influence of the unconnected 3D simplexes describing axial chirality can be caused by the fact that pairs of atoms in such simplexes are

Table 12 Structures and antiplasmodial activity of the studied naphthylisoquinoline alkaloids

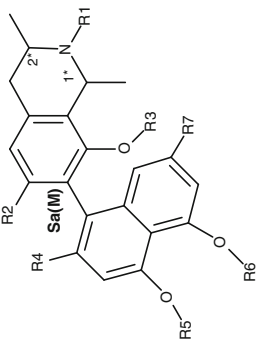
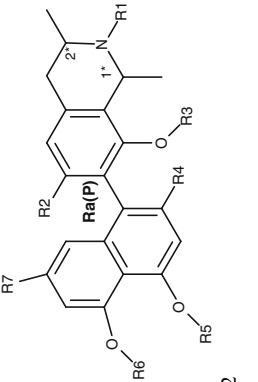
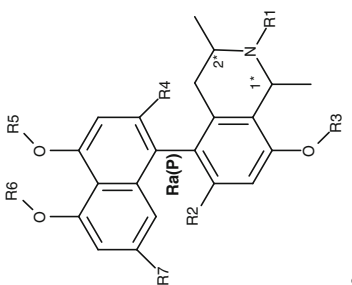
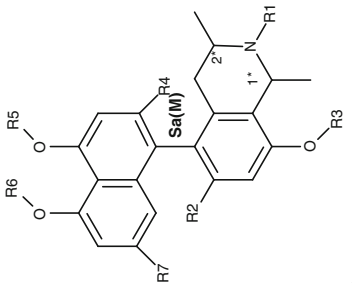
 <p style="text-align: center;">1</p>	 <p style="text-align: center;">2</p>									
 <p style="text-align: center;">3</p>	 <p style="text-align: center;">4</p>									
<p>I* and 2* are labels of chiral centers and R_a(P) and S_a(M) are configurations of the corresponding axis of chirality</p>										
Compound	R1	R2	R3	R4	R5	R6	R7	1* ⁱ	2* ⁱ	logC ₅₀
1a	H	H	H	CH ₂ OH	CH ₃	H	H	R	R	1.88
1b	a	OCH ₃	CH ₃	CH ₃	CH ₃	CH ₃	H	a	S	-0.236
(continued)										

Table 12 (continued)

1c	H	OCH ₃	CH ₃	CH ₃	CH ₃	CH ₃	CH ₃	H	R	S	-0.237
1d	H	OCH ₃	CH ₃	CH ₃	CH ₃	CH ₃	CH ₃	H	S	S	-0.265
1e	H	OH	CH ₃	CH ₃	CH ₃	CH ₃	CH ₃	H	R	R	-0.687
1f	H	OH	CH ₃	H	CH ₃	CH ₃	CH ₃	CH ₃	S	S	0.729
1g	H	OH	CH ₃	CH ₃	CH ₃	CH ₃	CH ₃	H	R	S	-0.639
1h	CH ₂ C ₆ H ₅	H	H	CH ₂ OH	CH ₃	CH ₃	CH ₃	H	R	R	0.299
1i	CH ₂ C ₆ H ₅	H	CH ₂ C ₆ H ₅	CH ₂ OH	CH ₃	CH ₃	CH ₃	H	R	R	-0.001
1j	H	H	H	CH ₂ OH	CH ₃	CH ₃	CH ₃	H	R	R	1.896
1k	H	H	H	CH ₃	CH ₃	CH ₃	CH ₃	H	R	R	0.419
1l	H	H	H	CH ₃	CH ₃	CH ₃	CH ₃	H	R	R	0.381
1m	H	H	H	CH ₃	H	H	H	H	R	R	-0.343
2a	H	H	H	CH ₂ OH	CH ₃	CH ₃	CH ₃	H	R	R	0.538
2b	H	OCH ₃	H	H	H	CH ₃	CH ₃	CH ₃	S	S	-0.138
2c	CH ₂ C ₆ H ₅	H	H	CH ₂ OH	CH ₃	CH ₃	CH ₃	H	R	R	0.257
2d	H	H	H	CH ₃	CH ₃	CH ₃	CH ₃	H	R	R	0.516
2e	CH ₃	H	H	CH ₃	CH ₃	CH ₃	CH ₃	H	R	R	-0.172
2f	H	H	H	H	H	CH ₃	CH ₃	CH ₃	R	R	0.797
2g	CH ₃	OCH ₃	CH ₃	H	CH ₃	CH ₃	CH ₃	CH ₃	S	S	-0.369
3a	H	H	H	CH ₃	CH ₃	CH ₃	CH ₃	H	R	R	1.782
3b	a	H	CH ₃	CH ₃	CH ₃	CH ₃	CH ₃	H	a	b	-0.261
3c	CH ₃	OH	CH ₃	H	CH ₃	CH ₃	CH ₃	CH ₃	R	R	-0.853
3d	H	OH	CH ₃	H	CH ₃	CH ₃	CH ₃	H	S	R	0.226
3e	CH ₃	OH	CH ₃	H	CH ₃	CH ₃	CH ₃	CH ₃	R	R	-0.356
3f	CH ₃	OCH ₃	CH ₃	CH ₃	CH ₃	CH ₃	CH ₃	CH ₃	R	S	-0.868
3g	CH ₃	H	H	CH ₃	CH ₃	CH ₃	CH ₃	H	R	R	0.453

(continued)

Table 12 (continued)

3h	H	OH	H	H	H	H	CH ₃	CH ₃	CH ₃	R	R	0.936
4a	H	H	CH ₃	CH ₃	CH ₃	CH ₃	CH ₃	CH ₃	CH ₃	S	S	-0.542
4b	CH ₃	H	CH ₃	CH ₃	CH ₃	CH ₃	CH ₃	CH ₃	CH ₃	S	S	-0.329
4c	CH ₃	H	H	H	H	H	CH ₃	CH ₃	CH ₃	R	R	0.3
4d	H	OCH ₃	H	H	CH ₃	CH ₃	CH ₃	CH ₃	CH ₃	R	R	0.423
4e	a	OCH ₃	CH ₃	H	H	CH ₃	CH ₃	CH ₃	CH ₃	A	S	-0.401
4f	H	OCH ₃	CH ₃	H	H	H	CH ₃	CH ₃	CH ₃	S	S	-0.102
4g	CH ₃	OCH ₃	H	H	H	H	CH ₃	CH ₃	CH ₃	R	R	0.464
4h	a	H	CH ₃	H	CH ₃	CH ₃	H	CH ₃	CH ₃	a	R	-0.305
4i	CH ₃	H	CH ₃	H	H	CH ₃	CH ₃	CH ₃	CH ₃	S	S	-1.051
4j	CH ₃	CH ₃	CH ₃	H	H	CH ₃	CH ₃	CH ₃	CH ₃	R	S	-0.347
4k	a	H	CH ₃	CH ₃	CH ₃	CH ₃	CH ₃	H	H	a	b	-0.277
4l	H	H	H	H	H	H	CH ₃	CH ₃	CH ₃	R	R	0.738

ⁱa means the double bond between 1* and N in isoquinoline cycle; b means that the isoquinoline part is fully unsaturated

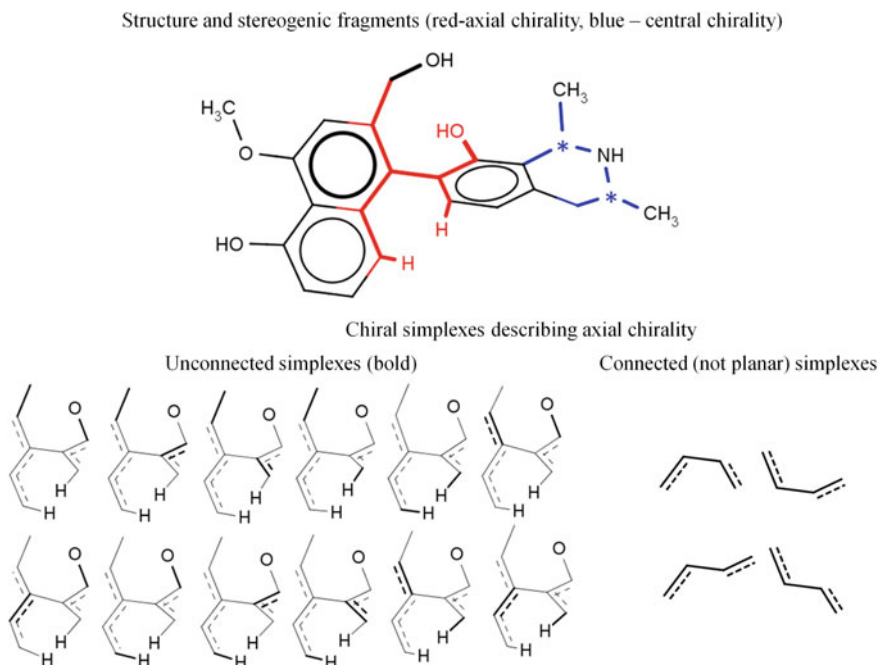


Fig. 23 The structure of dioncopeltine A and simplexes representing axial chirality of the compound

sufficiently separated in space (especially compared to connected simplexes). Thus, unconnected chiral simplexes can better describe the shape of a molecule in terms of its stereochemical features.

The use of $(2 + 0.X)D$ -QSAR approach based on the SiRMS opens a lot of opportunities to get important information about structural and stereochemical factors which allows to include all necessary information about stereochemistry of studied compounds without sampling of conformers.

8 Software Implementation of Structural and Physicochemical Interpretation of QSAR Models Based on SiRMS

Structural and physicochemical interpretation approaches were implemented in an open-source tool for knowledge mining of chemical data sets—SPCI (Polishchuk et al. 2016). The overall procedure is straightforward: SDF-file → automatic model building and validation → calculation of desired fragments' contributions → visualization of contributions. There exist two interpretation modes: structural interpretation only and structural and physicochemical interpretation.

Different simplex labeling schemes are used for them. In the first case, vertices in simplexes are labeled by an element, in the second case they are labeled according to the value of partial atomic charge, lipophilicity, H-bonding ability and refraction to represent electrostatic, hydrophobic, H-bonding and dispersive factors, correspondingly. Therefore, overall contribution calculated in the second case will slightly differ from a contribution calculated in structural interpretation mode, because different descriptor labeling schemes are used.

Modeling and validation are performed automatically, and no variable selection is performed during modeling. Optimal model parameters are tuned by a grid search. five-fold cross-validation is performed only once (one repetition) with the predefined seed to make it reproducible. Statistics and parameters used for optimal models building can be viewed in a separate window.

The exploratory analysis can be rapidly and easily performed for any data set based on several pre-defined fragmentation schemes: (1) common functional groups and small rings; (2) all ring systems available in the modeling data set; (3) Murcko scaffolds detected in the modeling data set; (4) two automatic fragmentation schemes, which use SMARTS to define bonds to cleave during fragmentation. In the latter fragmentation scheme, only fragments with at most three attachment points are created to avoid combinatorial explosion. The user-defined fragments in SMARTS/SMILES format may also be used for fragmentation (tab-separated list of SMART/SMILES and their names). There are lots of command line options in underlying Python scripts which provide great customization and tuning of the whole system. Though many parameters are set to the reasonable default values to simplify the usage and graphical interface. In the last 0.1.5 version the predictor module was added, which returns predictions for new data sets and estimates the applicability domain based on a fragment control approach. Only basic visualization of structure-activity relationship trends was implemented in SPCI software. More advanced and flexible visualization is provided by rspci R package (<https://github.com/DrrDom/rspci>).

For more details on the software tools, one can refer to the SPCI manual and the author page http://qsar4u.com/pages/sirms_qsar.php. The open-source code is available in the following github repositories: (i) standalone SPCI software tool with GUI (<https://github.com/DrrDom/spci>); (ii) Python tool to carry out fragmentation of the data set compounds for further descriptors calculation with external software and prediction by QSAR models and for calculation of fragment contributions (<https://github.com/DrrDom/spci-ext>); (iii) R package for customized visualization of fragment contributions (<https://github.com/DrrDom/rspci>).

9 Conclusions

The simplex representation of molecular structure proved itself as a powerful and flexible tool for encoding chemical structures in QSAR modeling. The SiRMS allows not only to represent topology of molecules but their stereochemical configuration regardless their type of chirality. The simplex descriptors are suitable for an interpretation of QSAR models by means of model-specific and model-independent approaches. We demonstrated that one should consider three different scenarios of interpretation based on a mechanism of action of studied compounds for reasonable interpretation results. In all case studies, we obtained relevant and reasonable interpretation results. Models showed high performance, and thus the SiRMS descriptors can be recommended to build highly predictive and interpretable QSAR models.

References

- Adenot, M., & Lahana, R. (2004). Blood-brain barrier permeation models: Discriminating between potential CNS and non-CNS drugs including P-glycoprotein substrates. *Journal of Chemical Information and Computer Sciences*, 44(1), 239–248. doi:10.1021/ci034205d.
- Aires-de-Sousa, J., & Gasteiger, J. (2002). Prediction of enantiomeric selectivity in chromatography: Application of conformation-dependent and conformation-independent descriptors of molecular chirality. *Journal of Molecular Graphics and Modelling*, 20(5), 373–388. doi:10.1016/S1093-3263(01)00136-X.
- Andrieux, A., et al. (1989). Amino acid sequences in fibrinogen mediating its interaction with its platelet receptor, GPIIb/IIIa. *Journal of Biological Chemistry*, 264(16), 9258–9265.
- Avdeef, A. (2012). *Absorption and drug development: Solubility, permeability, and charge state* (2nd ed.). Hoboken, NJ: Wiley-Interscience.
- Beechey, R. B. (1966). The uncoupling of respiratory-chain phosphorylation by 4,5,6,7-tetrachloro-2-trifluoromethylbenzimidazole. *Biochemical Journal*, 98(1), 284–289.
- Besalú, E., Gironés, X., Amat, L., & Carbó-Dorca, R. (2002). Molecular quantum similarity and the fundamentals of QSAR. *Accounts of Chemical Research*, 35(5), 289–295. doi:10.1021/ar010048x.
- Bikadi, Z., et al. (2011). Predicting P-glycoprotein-mediated drug transport based on support vector machine and three-dimensional crystal structure of P-glycoprotein. *PLoS ONE*, 6(10), e25815.
- Bringmann, G., & Rummey, C. (2003). 3D QSAR investigations on antimalarial naphthylisoquinoline alkaloids by comparative molecular similarity indices analysis (CoMSIA), based on different alignment approaches. *Journal of Chemical Information and Computer Sciences*, 43(1), 304–316. doi:10.1021/ci025570s.
- Bukowska, B. (2006). Toxicity of 2,4-dichlorophenoxyacetic acid—Molecular mechanisms. *Polish Journal of Environmental Studies*, 15(3), 365–374.
- Carbonell, P., Carlsson, L., & Faulon, J.-L. (2013). Stereo signature molecular descriptor. *Journal of Chemical Information and Modeling*, 53(4), 887–897. doi:10.1021/ci300584r.
- Casarett, L. J., & Klaassen, C. D. (2008). *Casarett and Doull's toxicology: The basic science of poisons* (7th ed.). New York: McGraw-Hill Medical.
- Cherkasov, A., et al. (2014). QSAR modeling: Where have you been? Where are you going to? *Journal of Medicinal Chemistry*, 57(12), 4977–5010. doi:10.1021/jm4004285.

- Čolović, M. B., Krstić, D. Z., Lazarević-Pašti, T. D., Bondžić, A. M., & Vasić, V. M. (2013). Acetylcholinesterase inhibitors: Pharmacology and toxicology. *Current Neuropharmacology*, 11(3), 315–335. doi:10.2174/1570159x11311030006.
- Cramer, R. D., Patterson, D. E., & Bunce, J. D. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society*, 110(18), 5959–5967. doi:10.1021/ja00226a005.
- cxcalc. 5.4 edn. Chemaxon, Budapest, Hungary.
- Davidson, B., Soodak, M., Strout, H. V., Neary, J. T., Nakamura, C., & Maloof, F. (1979) Thiourea and cyanamide as inhibitors of thyroid peroxidase: The role of iodide. *Endocrinology*, 104(4), 919–924. doi:10.1210/endo-104-4-919.
- Dunn, J. F., Nisula, B. C., & Rodbard, D. (1981). Transport of steroid hormones: Binding of 21 endogenous steroids to both testosterone-binding globulin and corticosteroid-binding globulin in human plasma. *The Journal of Clinical Endocrinology & Metabolism*, 53(1), 58–68. doi:10.1210/jcem-53-1-58.
- Egbertson, M. S., Chang, C. T. C., Duggan, M. E., Gould, R. J., Halczenko, W., Hartman, G. D., et al. (1994). Non-peptide fibrinogen receptor antagonists. 2. Optimization of a tyrosine template as a mimic for Arg-Gly-Asp. *Journal of Medicinal Chemistry*, 37(16), 2537–2551. doi:10.1021/jm00042a007.
- Free, S. M., & Wilson, J. W. (1964). A mathematical contribution to structure-activity studies. *Journal of Medicinal Chemistry*, 7(4), 395–399. doi:10.1021/jm00334a001.
- Fujita, T., & Winkler, D. A. (2016). Understanding the roles of the “Two QSARs”. *Journal of Chemical Information and Modeling*, 56(2), 269–274. doi:10.1021/acs.jcim.5b00229.
- Gartner, T. K., & Bennett, J. S. (1985). The tetrapeptide analogue of the cell attachment site of fibronectin inhibits platelet aggregation and fibrinogen binding to activated platelets. *Journal of Biological Chemistry*, 260(22), 11891–11894.
- Gasteiger, J. (2016). Chemoinformatics: Achievements and challenges, a personal view. *Molecules*, 21(2), 151.
- Gaulton, A., et al. (2012). ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1), D1100–D1107. doi:10.1093/nar/gkr777
- Ghose, A. K., Herbertz, T., Hudkins, R. L., Dorsey, B. D., & Mallamo, J. P. (2012). Knowledge-based, central nervous system (CNS) lead selection and lead optimization for CNS drug discovery. *ACS Chemical Neuroscience*, 3(1), 50–68. doi:10.1021/cn200100h.
- Golbraikh, A., Bonchev, D., & Tropsha, A. (2001). Novel chirality descriptors derived from molecular topology. *Journal of Chemical Information and Computer Sciences*, 41(1), 147–158. doi:10.1021/ci000082a.
- Grundlingh, J., Dargan, P., El-Zanfaly, M., & Wood, D. (2011). 2,4-Dinitrophenol (DNP): A weight loss agent with significant acute toxicity and risk of death. *Journal of Medical Toxicology*, 7(3), 205–212. doi:10.1007/s13181-011-0162-6.
- Guha, R. (2008). On the interpretation and interpretability of quantitative structure–activity relationship models. *Journal of Computer-Aided Molecular Design*, 22(12), 857–871. doi:10.1007/s10822-008-9240-5.
- Hammitt, L. P. (1937). The effect of structure upon the reactions of organic compounds. Benzene derivatives. *Journal of the American Chemical Society*, 59(1), 96–103. doi:10.1021/ja01280a022.
- Hartman, G. D., et al. (1992). Non-peptide fibrinogen receptor antagonists. 1. Discovery and design of exosite inhibitors. *Journal of Medicinal Chemistry*, 35(24), 4640–4642. doi:10.1021/jm00102a020.
- Hitchcock, S. A. (2012). Structural modifications that alter the P-glycoprotein efflux properties of compounds. *Journal of Medicinal Chemistry*, 55(11), 4877–4895. doi:10.1021/jm201136z.
- Hitchcock, S. A., & Pennington, L. D. (2006). Structure—Brain exposure relationships. *Journal of Medicinal Chemistry*, 49(26), 7559–7583. doi:10.1021/jm060642i.
- <http://www2.epa.gov/chemical-research/toxicity-estimation-software-tool-test>.

- Joback, K. G., & Reid, R. C. (1987). Estimation of pure-component properties from group-contributions. *Chemical Engineering Communications*, 57(1–6), 233–243. doi:10.1080/00986448708960487.
- Kansy, M., Senner, F., & Gubernator, K. (1998). Physicochemical high throughput screening: Parallel artificial membrane permeation assay in the description of passive absorption processes. *Journal of Medicinal Chemistry*, 41(7), 1007–1010. doi:10.1021/jm970530e.
- Kazius, J., McGuire, R., & Bursi, R. (2005). Derivation and validation of toxicophores for mutagenicity prediction. *Journal of Medicinal Chemistry*, 48(1), 312–320. doi:10.1021/jm040835a.
- Kuz'min, V. E., Artemenko, A. G., & Muratov, E. N. (2008). Hierarchical QSAR technology based on the simplex representation of molecular structure. *The Journal of Computer-Aided Molecular Design*, 22(6–7), 403–421. doi:10.1007/s10822-008-9179-6.
- Kuz'min, V. E., et al. (2005). Hierarchic system of QSAR models (1D-4D) on the base of simplex representation of molecular structure. *Journal of Molecular Modelling*, 11, 457–467.
- Leach, A. G., et al. (2006). Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *Journal of Medicinal Chemistry*, 49(23), 6672–6682. doi:10.1021/jm0605233.
- Lindberg, B., Svensson, S., Malmquist, P. Å., Basilier, E., Gelius, U., & Siegbahn, K. (1976). Correlation of ESCA shifts and Hammett substituent constants in substituted benzene derivatives. *Chemical Physics Letters*, 40(2), 175–179. doi:10.1016/0009-2614(76)85053-1.
- Littin, K. E., O'Connor, C. E., & Eason, C. T. (2000). Comparative effects of brodifacoum on rats and possums. *New Zealand Plant Protection*, 53, 310–315.
- Liu, S.-S., Yin, C.-S., & Wang, L.-S. (2002). Combined MEDV-GA-MLR method for QSAR of three panels of steroids, dipeptides, and COX-2 inhibitors. *Journal of Chemical Information and Computer Sciences*, 42(3), 749–756. doi:10.1021/ci010245a.
- Lobato, M., Amat, L., Besalú, E., & Carbó-Dorca, R. (1997). Structure-activity relationships of a steroid family using quantum similarity measures and topological quantum similarity indices. *Quantitative Structure-Activity Relationships*, 16(6), 465–472. doi:10.1002/qsar.19970160605.
- Lukovits, I., & Linert, W. (2001). A topological account of chirality. *Journal of Chemical Information and Computer Sciences*, 41(6), 1517–1520. doi:10.1021/ci0100346.
- Ma, X.-L., Chen, C., & Yang, J. (2005). Predictive model of blood-brain barrier penetration of organic compounds. *Acta Pharmacologica Sinica*, 26(4), 500–512.
- Marrero-Morejón, J., & Pardillo-Fontdevila, E. (1999). Estimation of pure compound properties using group-interaction contributions. *AIChE Journal*, 45(3), 615–621. doi:10.1002/aic.690450318.
- Marrero-Ponce, Y., Castillo-Garit, J. A., Castro, E. A., Torrens, F., & Rotondo, R. (2008). 3D-chiral (2.5) atom-based TOMOCOMD-CARDD descriptors: Theory and QSAR applications to central chirality codification. *Journal of Mathematical Chemistry*, 44(3), 755–786. doi:10.1007/s10910-008-9386-3.
- Medina, M. A. (1963). The in vivo effects of hydrazines and vitamin B6 on the metabolism of gamma-aminobutyric acid. *Journal of Pharmacology and Experimental Therapeutics*, 140(2), 133–137.
- O'Brien, R. D., Kirkpatrick, M., & Miller, P. S. (1964). Poisoning of the rat by hydrazine and alkylhydrazines. *Toxicology and Applied Pharmacology*, 6(4), 371–377. doi:10.1016/S0041-008X(64)80001-6.
- Parretti, M. F., Kroemer, R. T., Rothman, J. H., & Richards, W. G. (1997). Alignment of molecules by the Monte Carlo optimization of molecular similarity indices. *Journal of Computational Chemistry*, 18(11), 1344–1353. doi:10.1002/(sici)1096-987x(199708)18:11<1344:aid-jcc2>3.0.co;2-1.
- Polishchuk, P., et al. (2016). Structural and physico-chemical interpretation (SPCI) of QSAR models and its comparison with matched molecular pair analysis. *Journal of Chemical Information and Modeling*. doi:10.1021/acs.jcim.6b00371.

- Polishchuk, P. G., Kuz'min, V. E., Artemenko, A. G., & Muratov, E. N. (2013). Universal approach for structural interpretation of QSAR/QSPR models. *Molecular Informatics*, 32(9–10), 843–853. doi:10.1002/minf.201300029.
- Polishchuk, P. G., et al. (2015). Design, virtual screening, and synthesis of antagonists of α Ib β 3 as antiplatelet agents. *Journal of Medicinal Chemistry*, 58(19), 7681–7694. doi:10.1021/acs.jmedchem.5b00865.
- Proudfoot, A., Bradberry, S., & Vale, J. A. (2006). Sodium fluoroacetate poisoning. *Toxicological Reviews*, 25(4), 213–219. doi:10.2165/00139709-200625040-00002.
- Reid, R. C., Prausnitz, J. M., & Poling, B. E. (1987). *The properties of gases and liquids* (4th ed.). New York: McGraw-Hill.
- Riniker, S., & Landrum, G. (2013). Similarity maps—A visualization strategy for molecular fingerprints and machine-learning methods. *Journal of Cheminformatics*, 5(1), 43.
- Scarborough, R. M., et al. (1993). Design of potent and specific integrin antagonists. Peptide antagonists with high specificity for glycoprotein IIb-IIIa. *Journal of Biological Chemistry*, 268(2), 1066–1073.
- Silverman, B. D., & Platt, D. E. (1996). Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. *Journal of Medicinal Chemistry*, 39(11), 2129–2140. doi:10.1021/jm950589q.
- Springer, T. A., Zhu, J., & Xiao, T. (2008). Structural basis for distinctive recognition of fibrinogen γ C peptide by the platelet integrin α Ib β 3. *The Journal of Cell Biology*, 182(4), 791–800. doi:10.1083/jcb.200801146.
- Sushko, Y., Novotarskyi, S., Korner, R., Vogt, J., Abdelaziz, A., & Tetko, I. (2014). Prediction-driven matched molecular pairs to interpret QSARs and aid the molecular optimization process. *Journal of Cheminformatics*, 6(1), 48.
- Takahata, Y., & Chong, D. P. (2005). Estimation of Hammett sigma constants of substituted benzenes through accurate density-functional calculation of core-electron binding energy shifts. *International Journal of Quantum Chemistry*, 103(5), 509–515. doi:10.1002/qua.20533.
- Terada, H. (1990). Uncouplers of oxidative phosphorylation. *Environmental Health Perspectives*, 87, 213–218.
- Thermodynamics Research Center, NIST Boulder Laboratories, M. Frenkel director (2013). Thermodynamics source database. In P. J. Linstrom & W. G. Mallard (Eds.), NIST chemistry WebBook, NIST standard reference database number 69. National Institute of Standards and Technology, Gaithersburg MD, 20899.
- Valchev, I., Binev, R., Yordanova, V., & Nikolov, Y. (2008). Anticoagulant rodenticide intoxication in animals—A review. *Turkish Journal of Veterinary and Animal*, 32(4), 237–243.
- Wager, T. T., et al. (2010a). Defining desirable central nervous system drug space through the alignment of molecular properties, in vitro ADME, and safety attributes. *ACS Chemical Neuroscience*, 1(6), 420–434. doi:10.1021/cn100007x.
- Wager, T. T., Hou, X., Verhoest, P. R., & Villalobos, A. (2010b). Moving beyond rules: The development of a central nervous system multiparameter optimization (CNS MPO) approach to enable alignment of drug like properties. *ACS Chemical Neuroscience*, 1(6), 435–449. doi:10.1021/cn100008c.
- Wassermann, A. M., Haebel, P., Weskamp, N., & Bajorath, J. (2012). SAR matrices: Automated extraction of information-rich SAR tables from large compound data sets. *Journal of Chemical Information and Modeling*, 52(7), 1769–1776. doi:10.1021/ci300206e.
- Weininger, D., Weininger, A., & Weininger, J. L. (1989). SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences*, 29(2), 97–101. doi:10.1021/ci00062a008.

The Maximum Common Substructure (MCS) Search as a New Tool for SAR and QSAR

Azadi Golbamaki, Alessio Mauro Franchi and Giuseppina Gini

Abstract The Maximum Common Substructure (MCS) between two molecules induces a similarity that makes it possible to group compounds sharing the same pattern. In our study the relevance of a similarity measure exclusively based on MCS has been implemented in new software based on the `fmcs_R` package. The newly developed program searches for the largest substructures between a target molecule, with unknown property value, and a set of similar molecules with experimental value to assess the toxicity of the target chemical. In QSAR and read-across, while reasoning on the similarity of the evaluated molecules, another important aspect to consider is the difference of two molecules that share a large common part. Thus, the present study examines the issue of the MCS itself, and the differences between a reference and a similar molecule by the aid of an ad hoc developed software. The most important features of this software are: (I) the process of the MCSs between two molecules represented as graphs and (II) the detection and the graphical representation of the dissimilar substructures that are identified in the target and the source molecules. The user may consequently quantify the properties and weights of these substructures to improve the assessment of new substances. This new software is integrated into ToxRead, a system to visualize structures and substructures for expert reasoning. Moreover, an automatic search in a database containing the role of small substructures in amplifying or reducing the property can help in improving the final assessment.

Keywords Graph theory • Read-across • Chemical similarity • Structural alerts

A. Golbamaki (✉)

Istituto di Ricerche Farmacologiche “Mario Negri” Milano, Milan, Italy
e-mail: azadi.golbamaki@marionegri.it

A.M. Franchi • G. Gini

DEIB, Politecnico di Milano, Milan, Italy
e-mail: alessiomauro.franchi@polimi.it

G. Gini

e-mail: giuseppina.gini@polimi.it

© Springer International Publishing AG 2017

K. Roy (ed.), *Advances in QSAR Modeling*, Challenges and Advances

in Computational Chemistry and Physics 24, DOI 10.1007/978-3-319-56850-8_5

1 Introduction

Clustering and classification methods used to group similar chemicals could benefit from the Maximum Common Substructure (MCS) algorithm. The MCS algorithm usually takes two molecular structures, represented as graphs, and extracts the maximum common substructure, either as a connected or disconnected graph. Usual applications of MCS are for filtering and prioritizing large data sets of molecules. Structure-Activity Relationship (SAR) systems often search for large substructures with known properties to predict the property of new molecules that contain them; this task is substructure search, a much simpler case than MCS.

In Quantitative Structure-Activity Relationship (QSAR) and read-across, while reasoning on the similarity of the considered molecules, another need may arise: how to account for the differences of two molecules that share a large common part? What is missing here is not the maximum common substructure itself, but exactly what are the differences between a reference and a test molecule. In this chapter we discuss how to detect and show these differences, how to quantify their weight, and how to use them to improve the assessment of new substances.

Finding the maximum common substructure is a computationally hard algorithm that has been studied in literature and has received many complete or heuristic solutions (Garey and Johnson 1979). The MCS problem is recognized as a “NP” problem (for which no polynomial-time solutions are known), and complete solutions are indeed exponential in the number of atoms of the substructure. We review the main algorithms for MCS and discuss our specific approach tailored for our aims.

2 Notation, Basic Algorithm and Classical Applications of MCS

Determining a one-to-one atom correspondence between two chemical compounds is important to measure molecular similarities. This determination is a case of well-known problems on graphs.

2.1 Basic Definitions

A **graph** is a collection of n vertices and m edges connecting them; formally $G = (V, E)$, where E is a set of unordered pairs from V . In undirected graphs the edges have no orientation.

The **subgraph isomorphism** problem is finding a fixed graph as a subgraph in a given graph. A particular subgraph is a clique.

A **clique** in a $G = (V, E)$ is a subset $S \subseteq V$ of vertices of an undirected graph such that the induced subgraph is complete, i.e. every two distinct vertices in the clique are adjacent. Mathematically we define a clique as a subset of a directed graph satisfying the following conditions:

- The subset contains at least three points.
- If P_i and P_j are in the clique, then there is an edge connecting P_i and P_j .
- The subset is the largest possible.

A **maximal clique** is a clique that cannot be extended by including one more adjacent vertex. A graph with $3n$ vertices can have at most 3^n maximal cliques, as demonstrated by Moon and Moser (1965).

The **maximum common subgraph-isomorphism (MCS)** problem is defined as a decision problem: given two graphs, G_1 and G_2 , what is the largest subgraph of G_1 isomorphic to a subgraph of G_2 ? MCS is NP-hard too.

2.2 Computational Approaches to Clique

The computational problem of finding (all) the largest complete subgraph(s) (maximal clique) is called the clique problem. Since their number can be very large, smart algorithms need to be designed. The clique problem is NP-complete, i.e. no polynomial time algorithms have been found to solve the general problem. Many algorithms for computing cliques have been developed, both complete and approximate.

A well-known complete algorithm is by Bron and Kerbosh (1973); it finds all cliques, in an undirected graph, running in exponential time. Ullmann (1976) developed a subgraph isomorphism method, applied also to clique detection, and optimized through special hardware; it significantly reduces the size of the search space using backtracking.

Bunke and Messmer (1995) attempted to reduce the overall computational cost, resulting in a quadratic time with respect to graph size, but with an exponential memory requirement and pre-processing time. Other techniques, such as non-deterministic ones, reduce the complexity from exponential to polynomial, but are not guaranteed to find an exact and optimal solution. Cordella et al. (2004) worked to reduce the memory needs so as to scale the system to thousands of nodes and branches. Finally, to address the matching of very large graphs, Zhu et al. (2013) proposed an approximate solution with polynomial time complexity.

2.3 Use in Chemistry of Clique and MCS

In chemistry cliques and MCS algorithms are mostly used to describe chemical similarity. Chemical graphs are usually small graphs, with tens of nodes, but the chemical space is very large, with possibly millions of molecules (Reymond and Awale 2012).

Kuhl et al. (1983) used cliques to model the positions in which two chemicals will bind to each other. Gini et al. (2001) used the Ulmann algorithm to match expert-defined substructures related to carcinogenicity to tested chemicals. Cliques can be used to check a chemical dataset against a target structure, as in CLIP by Rhodes et al. (2003), that used the Bron-Kerbosch clique detection algorithm to find those structures in a file that have large structures in common with a target structure.

Raymond and Willett (2002) provided a classification and a review of the many MCS algorithms, both exact and approximate, which have been described in cheminformatics.

For MCS, typically, the number of bonds in the MCS is used as a similarity coefficient.

Cuissart et al. (2002) explored the use of MCS to define structural similarity indices and then to classify biodegradability of chemicals. Duesbury et al. (2015) explored the MCS similarity against the fingerprints approach. A combination of fingerprints and MCS is also used by Stah et al. (2005) to cluster a large chemical database for applications as identifying the most frequently occurring scaffolds, selecting analogues, and in the prioritization of chemical libraries. Cao et al. (2008) compared an MCS algorithm to global similarity measurements, and used them to predict and cluster biologically active compounds.

In the RASCAL system Raymod et al. (2002) developed the maximum common edge subgraph (MCES) algorithm to calculate graph similarity; the algorithm is based on maximum clique.

Despite the fact that most of the authors approach only one problem at a time, Xu (1996) has shown in his GMA algorithm that homomorphism, isomorphism, and maximal common substructure match (MCSS) can be processed in one algorithm with a complexity that depends on the number of edges.

Commercial systems are also integrating some MCS method, as in ChemAxon (Englert and Kovacs 2015). Free software is available; see for example the fMCS algorithm in python on <https://bitbucket.org/dalke/fmcs>.

3 Improving Read-Across Methods with Automatic Notification of Differences

Similarity of chemical structures plays an important role in grouping chemicals for read-across methods. Read-across assumes that a property or activity of a molecule depends on its chemical structure (Hansch and Leo 1979). We can assign a new

Table 1 Read-across, interpolation, and extrapolation of activity values from experimental values available for similar molecules in a group

	Chemical	Chemical _{i+1}	Chemical _{i+2}	Chemical _{i+3}	
activity1	Exp ⇒	Filled	Exp ⇒	Filled	<i>Read-across</i>
activity2	Exp ⇒	Filled	Filled ⇐	Exp	<i>Interpolation</i>
activity3	Filled ⇐	Exp	Exp ⇒	Filled	<i>Extrapolation</i>

molecule with missing values to a group of chemicals with high similarity and use the experimental values available in the group to induce the value of the new molecule. This way of reasoning also assumes that the values of a property have a kind of local linear behaviour; so it is necessary limiting the search area to the most similar molecules.

According to the Organization for Economic Cooperation and Development (OECD) documents,¹ this kind of read-across to fill missing values can be done considering the chemical group as a 2D matrix, where rows are properties and columns are molecules, as indicated in Table 1. The unknown property values can be filled by considering just one very similar molecule or two molecules to make interpolation or extrapolation.

Read-across relies on the expert experience in choosing the most similar molecule and the meaning of the functional subgroups in it. Several issues have been found; for instance, the data filled are not easily reproducible, as indicated in Benfenati et al. (2016), and the method itself is unable to make use of any statistical knowledge about the considered group of chemicals.

Of course the data filling can be done using SAR or statistical QSAR systems that usually work on molecules in a broader chemical space. However, there are problems also with them. In many cases regulators do not accept the results of QSAR, saying that they are not transparent. SAR methods, that assign the new molecule to the toxicity class in case it contains a known functional subgroup (structural alert), tend to overestimate the toxicity: in fact, the entire group of chemicals containing the structural alert is considered as toxic. This is the case, for instance, of the structural alerts for mutagenicity used in the ToxTree² software; their presence in the data set originally used to find them never reaches 100%, and in some cases is lower than 50%.

To improve the quality of QSAR models many solutions are available today, such as mixing different levels of description. For instance, it is well known that local considerations about the presence of specific atoms can improve the predictivity of QSAR models based on global descriptors (Toropov et al. 2010). The idea of dissecting the molecule, taking its subparts, and reasoning on them can be further exploited.

¹<http://www.oecd.org/chemicalsafety/risk-assessment/groupingofchemicalschemicalcategoriesandread-across.htm>.

²<http://www.toxtree.sourceforge.net/>.

To improve both the acceptability of QSAR and the reproducibility of read-across, we claim that it is worth reasoning also on the dissimilarities of molecules after considering their similarity. The reason is that the toxic activity may depend on the common part of the molecules, especially in case it contains a structural alert, but the remaining part can play a role in increasing or reducing the toxicity too.

Expert users are responsible, in read-across, in making this kind of considerations; however, the number of subgroups known to modify toxicity can be large and not fully memorized by the expert. Our proposed answer is a three-steps method:

- Find the MCS of two molecules (the reference and the unknown) and graphically show the structural differences.
- Check in a knowledge base of functional subgroups whether there is a structural alert in the MCS or in the differences.
- Check in a knowledge base whether the differences contain a toxicity modulator group.

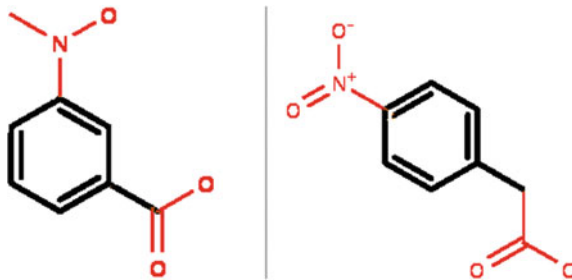
The following sections illustrate the first step and its implementation in the ToxRead (Gini et al. 2014) system. The other steps depend on the property under consideration and are in development for specific endpoints, starting from mutagenicity that is used in the following examples.

4 A New MCS Based Method to Detect Dissimilarity

As we have seen in the state of art, several solutions exist for computing the MCS, either optimal or not. The most common algorithms are focused on general graph theory; for example, several are based on converting the MCS problem into the clique problem, by introducing a compatibility graph. However, these methods do not exactly match with the representation needed for chemicals and consequently they cannot be very efficient. Moreover, the conversion to the clique problem prevents a flexible matching between two graphs. Stated these problems, along with the intractable computational complexity of the MCS problem, it is evident that more specific approaches are needed.

Our solution is based on the algorithm proposed in the `fmcs_R` package (Cao et al. 2008); it performs MCS computation via a novel backtracking algorithm, incrementally computing a search tree of correspondences between atoms of the two molecules under investigation. Each node in this tree is a set of atom correspondences, and leafs are the connected subgraphs we are looking for; the deepest leafs are the MCSs. This mechanism is more flexible than clique and makes it also possible to introduce various strategies to reduce the search space, such as pruning or branch and bound heuristics. It also let us define a mismatch tolerant comparison of atoms (i.e. different atoms may correspond for particular circumstances), using for example a priori defined set of admissible atom associations.

Fig. 1 The MCS in *black* is overlapped to both the molecule in study; the *red* branches are the desired differences



After the MCS is computed, we have to perform ring closure. Since the proposed algorithm does not consider rings as such, it may break some rings, i.e. it selects only a subset of atoms in a ring. This leads to a significant loss of structural information and consequently we need to close all the broken rings in the MCS.

We can finally extract the structural differences between the two compounds under investigation: we overlap each graph with the MCS and highlight all the sub-branches not in the MCS (Fig. 1).

4.1 The Algorithm

The backtracking algorithm works directly with the molecular graph structure; each atom of the compound is a vertex in the graph. The core idea is to search for all the possible combinations of two vertices correspondences, and build a search tree having sets of correspondences as nodes. As we move down the tree each set is possibly expanded with the addition of a new correspondence; each node is thus a subset of the set in each of its child nodes. The root of the tree is the empty set.

Figure 2 represents a simple example of a tree of correspondences, where A1, B1, C1, D1 and A2, B2, C2, D2 are nodes respectively from the first and the second molecular graph.

The matching between a node in the target graph and a second node in the query graph is driven by a set of rules, involving both the atoms itself (i.e. the atoms must be the equal) and their bonds (i.e. the two atoms must be connected by the same type of bonds). The expert may expand these rules by user-defined exceptions, possibly related to the particular molecules involved or the property it is trying to assess.

Once the tree is fully built, each leaf node is a candidate common subgraph, and the algorithm performs a depth-first search for the largest set of correspondences, i.e. the MCS. In the example of Fig. 2, one among the three lower nodes are returned; actually all the common subgraphs with maximum dimension are selected as the MCSs.

In order to reduce the complexity of the MCS problem and thus speed up the algorithm we used various strategies. The first concerns the selection of the next

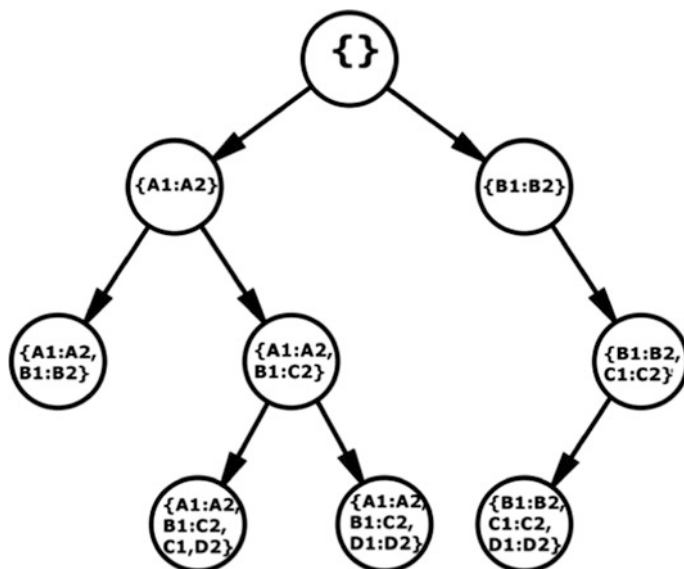


Fig. 2 A simple example of the search tree of correspondences. $A1, B1, C1, D1$ are atoms from the first molecule; $A2, B2, C2, D2$ are atoms from the second one

node in the target graph to be analysed; it is possible to create a sort of ordering in the search, with the aim of traversing as soon as possible branches that potentially contain an optimal solution. This allows the algorithm to prune out more branches and significantly reduce the dimension of the tree. In particular, we select the node having the highest number of neighbours in the current best common subgraph, resulting in a higher probability of findings a bigger subgraph.

The second strategy is related to the property of connectedness of the MCS. In general, a common subgraph may contain a number of disconnected fragments, but this is not desirable when working with chemical compounds. We then restrict our search to connected subgraph, greatly reducing the dimension of the search space. The implemented algorithm always expands the current common subgraph not creating disconnected fragments; for particular needs, it is also possible to relax this restriction introducing a maximum number of disconnected fragments. When this limit is reached during the traversal of a branch, or the current branch cannot be expanded further, a new branch starts.

4.2 Ring Closing

In the following processing, we have to close all the broken rings we have in the MCSs. With the term broken rings we are referring to those only partially included in the MCS; we can imagine the MCS as the longest path in common between two

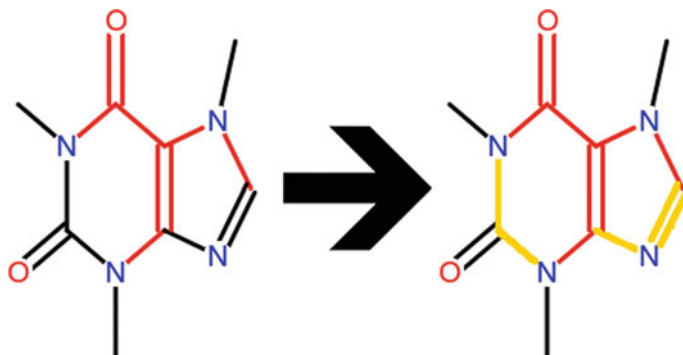


Fig. 3 An example of ring closure on the caffeine molecule; the MCS is highlighted in *red*; on the *left* there are two broken rings, which can be closed by the addition of the *yellow* atoms and bonds, as shown in the *right* graph

graphs and it may happen that during its computation it only partially traverse a ring, thus not preserving its structure. The left graph of Fig. 3 is an example: the MCS is highlighted in red and both the two rings are not fully in the MCS; these are two broken rings.

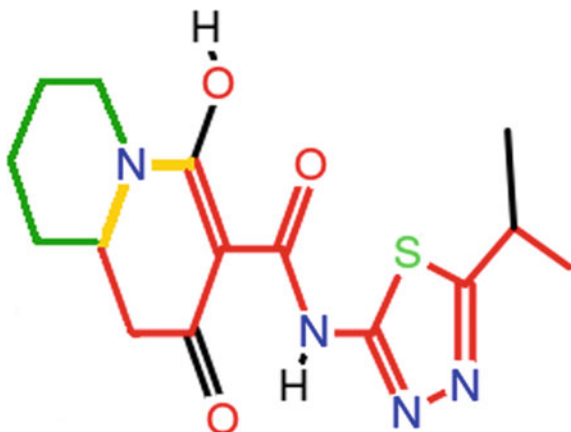
This step is fundamental for our goal because the information about the presence or not of a ring in the MCS (and consequently in the differences) often plays a major role in the final assessment of a compound property.

In order to close the rings, we need to keep track whether a node is in a ring or not; before we start computing the MCS, we analyse both the entire graphs and associate a positive flag with those nodes that are in a ring. At the same time, we also fill in a data structure containing a couple “ID - list of atoms” describing each ring in the graph. With these information, right after we extracted the MCS, we are able to close the ring: for each node in the MCS that has a positive flag, we manually add to the MCS all the other nodes that are in the same ring, but are still not present in the MCS. See right part of Fig. 3 for the final output of this step: red bonds and atoms are the MCS while yellow ones close the two rings.

It is important to state here that the updated subgraph is not anymore the exact MCS, as we relaxed it with the addition of several nodes; it is more a sort of “flexible-MCS”, but in our opinion this makes it is more adapt to capture a global and comprehensive information from the chemical side.

An interesting problem we faced with ring closure concerns double rings. The example in Fig. 4 clearly shows this issue. The molecule has an open ring (the second from the left) that we must close. There is also a second ring (the rightmost), which is already fully contained in the MCS and no further processing is required. Differently the green ring must not be added to the MCS as the only two nodes of it belonging to the MCS are shared with other rings. This particular case shows that the presence in the MCS of two atoms belonging to the same ring is not enough to

Fig. 4 A toy example with a non-real molecule; the MCS is drawn in *red*; *yellow* bonds must be added to close the broken ring, while *green* ones must not be selected to avoid a false positive ring



require a closure. To avoid these situations, we added a rule setting the minimum number of atoms of a ring included in the MCS required for its closure.

Even if ring closure may seem a quite time consuming procedure, it is infinitesimal with respect to the computational complexity of the MCS and its impact on the global performance is irrelevant.

4.3 Extracting the Differences

With the MCS ready, we extract the structural differences between the two compounds. The idea is to overlap the MCS with the molecules we are investigating, and select all the atoms and bonds not covered by the common part. The difficulty here resides in correctly detecting each connected subgraph.

We can practically do this in two separate steps. We first fill in the list of nodes and bonds (namely “nodeList” and “bondList”) of the target graph that have not been included in the MCS. If the MCS is equal to the graph, these two sets are empty and no residual subgraph exists. Secondly, we build each single connected branch one by one, in a recursive way: we start adding the first node of the first bond in the bondList to a new list, “nodeBranchList”. We now search for every bond in the bondList connected to that node, and add the second node of the bond to the nodeBranchList, also removing the bond from the bondList. We recursively repeat this step, each time selecting a new node to be expanded, until no more valid bonds are found. We now empty the nodeBranchList and repeat these procedures while the bondList is not empty. In this way, we build one tree of nodes for each connected branch not in the MCS. These are the structural differences we are looking for; Fig. 5 shows the three final trees we built starting from the bondList printed on the left. Each tree represents a connected subgraph.

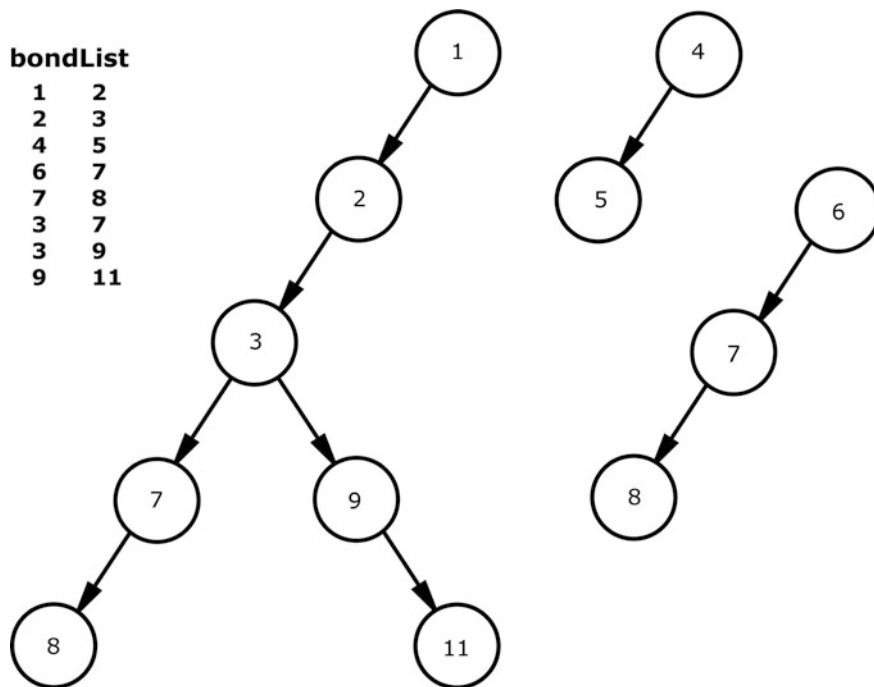


Fig. 5 The three connected subgraph built from the bond list shown on the *left*

5 Result and Discussion

Selecting the differences after applying the MCS is at the basis of the new functionality of ToxRead. As already mentioned, ToxRead is a free software to help experts in their read across activity by showing similar compounds and common alerts. A simple illustration of the ToxRead Graphical User Interface (GUI) is in Fig. 6.

The main window is a global view of the similar compounds, with the chemical under investigation drawn in the centre in light blue. Pop-up windows appear when clicking on molecules (circles) or rules (triangles). In particular, when clicking on the chemical, the following fields appear: chemical structure, CAS number, Similarity value, and Experimental value. When clicking on the rules, Chemical structure, Rule name, Rule description, and Rule accuracy appear.

In the newly introduced MCS-based dissimilarity function, the pop-up window shows the structures of the target and the similar molecules, the MCS, and the dissimilar substructures.

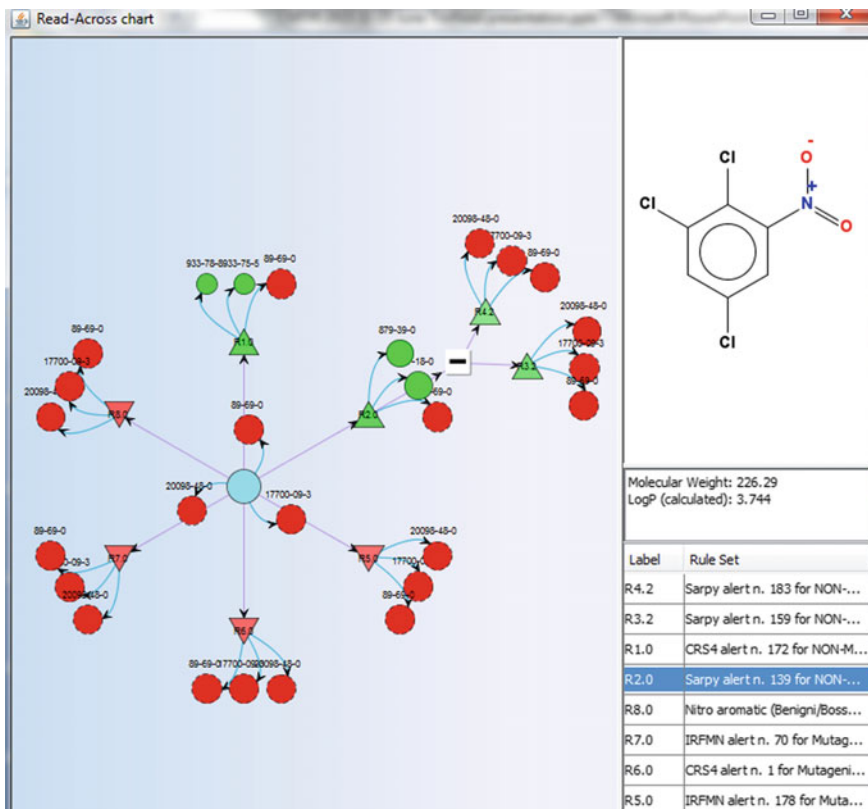


Fig. 6 The GUI of ToxRead. *Light blue circle* is the target compound; the others represent the most similar compounds

The addition of information about the structural differences between two similar compounds is the natural extension of the ToxRead software. In Fig. 7 we see the conceptual organization of the new functionality provided by MSC. As a new target compound is entered in the system, it is compared to the set of molecules in a dataset with known experimental values, and the most similar with respect to a similarity value are selected and displayed to the expert as before illustrated. The MCS and the dissimilarities between the target compound and a second similar molecule are computed as necessary. With the list of dissimilar substructures, the system now interrogates a dataset of rules, derived, for the specific endpoint, from a data set of compounds with experimental values. The subset of dissimilar substructures with toxicity information is eventually displayed to the user.

In the rest of this section we illustrate two simple case studies to show how the analysis of the differences can change the presumed straightforward classification of a molecule.

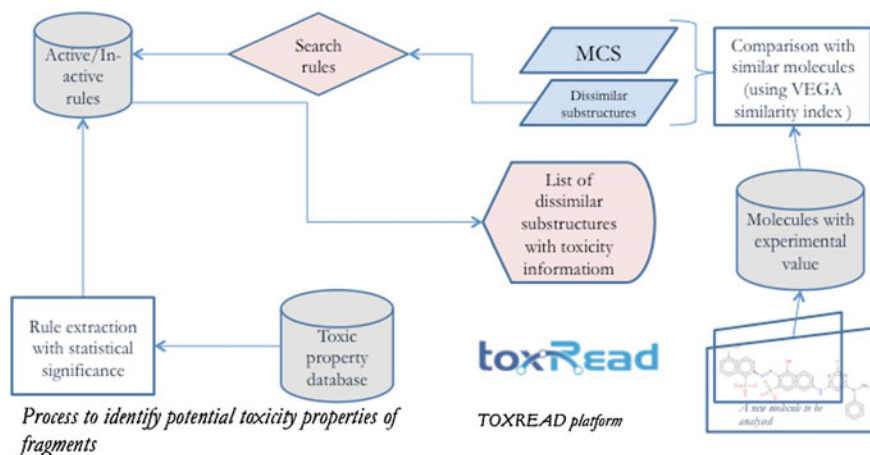


Fig. 7 The flow chart of the new dissimilarity system: on the *right* the use of MCS in ToxRead, on the *left* the interrogation of a database of specific rules for the fragments (obtained from an offline process)

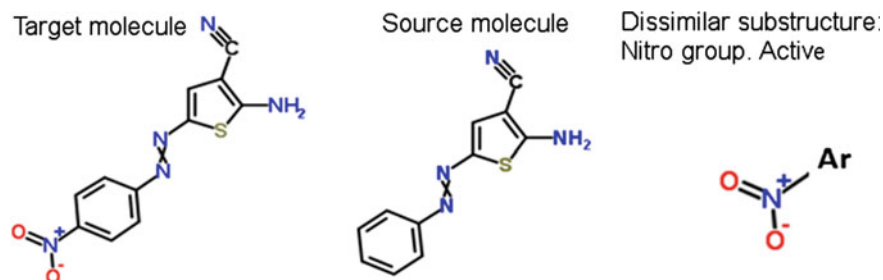


Fig. 8 The structural similarity and dissimilarity analysis between two molecules (Target molecule: 2-Amino-5-[(4-nitrophenyl)diazenyl]-3-thiophenecarbonitrile (unknown property); Source molecule: 2-Amino-5-(phenyldiazenyl)-3-thiophenecarbonitrile (property: non mutagen))

5.1 First Case Study

Let us consider two structurally similar molecules (their Tanimoto similarity index is 0.87 (Tanimoto 1958)) to investigate their mutagenicity property; suppose that the toxicity property of the target molecule was unknown.

We compared the target molecule with the source molecule with known mutagenicity property and identified the MCS and the dissimilar substructure by our software. Figure 8 shows the molecular structures, the MCS found, which is exactly the structure of the source molecule, and the dissimilar substructure.

The dissimilar fragment that is present in the target molecule is a nitride group substructure, which is known to be responsible for mutagenic property of a

molecule when it is connected to any aromatic ring in the molecule. In fact, the data on the experimental tests show that the target molecule is mutagenic. This simple example shows that the differences between two similar compounds can significantly change their property.

5.2 Second Case Study

Figure 9 shows another case study of two similar molecules (Tanimoto similarity index = 0.87) where the mutagenicity property of the target molecule is under investigation and the similar molecule is non mutagen. The MCS between two molecules has been identified, as well as the dissimilar substructures present in both molecules.

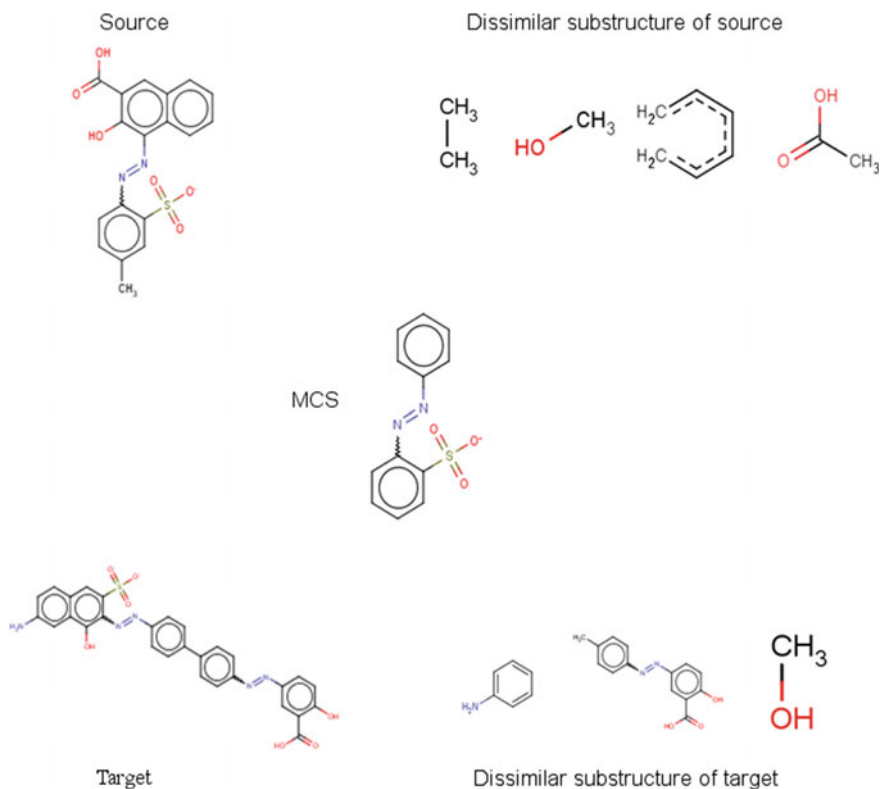


Fig. 9 The structural similarity and dissimilarity analysis between two molecules (Target molecule: 3-Hydroxy-4-[(4-methyl-2-sulfophenyl)diazenyl]-2-naphthoic acid; Source molecule: 5-[(E)-{4'-[(E)-(7-Amino-1-hydroxy-3-sulfonato-2-naphthyl)diazenyl]-4-biphenyl}] diazenyl]-2-hydroxybenzoate)

In particular, phenylamine is among the dissimilar substructures identified in the target molecule, and it is known to be an active mutagenic fragment in the collection of mutagenicity rules established at Istituto di Ricerche Farmacologiche Mario Negri (IRFMN) using automatic (Gini et al. 2013) and expert evaluations. Considering this rule will improve the assessment giving more evidence to the positive activity of the compound. Indeed, from experimental results, the target molecule is known to be mutagenic while a simple read-across with a similar molecule would suggest a negative activity.

6 Conclusions

The MCS and the dissimilarity searching are shown to be effective in the study and property prediction of chemical compounds. The differences between a new molecule and a similar molecule with experimental property value are important while reasoning on the activity of a new molecular structure.

Our MCS-based difference extraction method, incorporated into a new software tool, can help researchers in decision-making and property assessments of the molecular structures under investigation. It can be used for read across, where only local information about one or two similar molecules is used, or in assessing the prediction of QSAR results, or in refining the results of SAR systems that apply structural alerts.

The new tool is freely available inside ToxRead, and shows in the graphical interface the subparts of the molecule under investigation that are different from the reference similar molecule. The expert can investigate the role of those differences

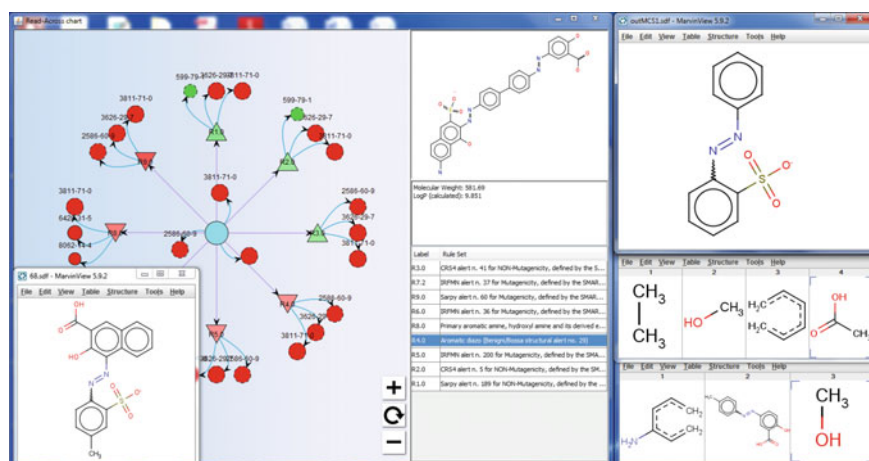


Fig. 10 An example of the integration of the new tool into ToxRead; this additional information may be used by the expert to improve its assessment of the chemical under investigation

on the basis of experience and knowledge. The next step will be the automatic search into a database of substructures relevant for the endpoint under study, so to alert the user about their presence. The construction of the knowledge base of substructures relevant for the mutagenicity endpoint is under development (Fig. 10).

Acknowledgements This research was supported by the PROSIL project (LIFE12 ENV/IT/000154). We thank Serena Manganelli and Giuseppa Raitano from the IRCCS—Istituto di Ricerca Farmacologiche Mario Negri, who provided insight and expertise that greatly assisted the research.

References

- Benfenati, E., Belli, M., Borges, T., Casimiro, E., Cester, J., Fernandez, A., et al. (2016). Results of a round-robin exercise on read-across. *SAR and QSAR in Environmental Research*, 27(5), 371–384. doi:10.1080/1062936X.2016.1178171.
- Bron, C., & Kerbosch, J. (1973). Finding all the cliques in an undirected graph. *Communication of the Association for Computing Machinery (ACM)*, 16(9), 189–201.
- Bunke, H., & Messmer, B. T. (1995). Efficient attributed graph matching and its application to image analysis. In *Proceeding of Image Analysis and Processing* (pp. 45–55). doi:10.1007/3-540-60298-4_235.
- Cao, Y., Jiang, T., & Girke, T. (2008). A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics*, 24(13), 366–374. doi:10.1093/bioinformatics/btn186.
- Cordella, L. P., Foggia, P., Sansone, C., & Vento, M. (2004). A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(10), 1367–1372.
- Cuissart, B., Touffet, F., Cremilleux, B., Bureau, R., & Rault, S. (2002). The maximum common substructure as a molecular depiction in a supervised classification context: Experiments in quantitative structure/biodegradability relationships. *Journal of Chemical Information and Modelling*, 42(5), 1043–1052. doi:10.1021/ci020017w.
- Duesbury, E., Holliday, J., & Willett, P. (2015). Maximum common substructure-based data fusion in similarity searching. *Journal of Chemical Information and Modelling*, 55(2), 222–230.
- Englert, P., & Kovacs, P. (2015). Efficient heuristics for maximum common substructure search. *Journal of Chemical Information and Modelling*, 55(5), 941–955.
- Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. : W.H. Freeman. ISBN 0-7167-1045-5.
- Gini, G., Ferrari, T., Cattaneo, D., Golbamaki, N., Manganaro, A., & Benfenati, E. (2013). Automatic knowledge extraction from chemical structures: The case of mutagenicity prediction. *SAR and QSAR in Environmental Research*, 24(5), 365–383.
- Gini, G., Franchi, A. M., Manganaro, A., Golbamaki, A., & Benfenati, E. (2014). ToxRead: A tool to assist in read across and its use to assess mutagenicity of chemicals. *SAR and QSAR in Environmental Research*, 25(12), 999–1011.
- Gini, G., Lorenzini, M., Benfenati, E., Brambilla, R., & Malvè, L. (2001). Mixing a symbolic and a subsymbolic expert to improve carcinogenicity prediction of aromatic compounds. In *Proceedings of the Second International Workshop on Multiple Classifier Systems (MCS 2001)*, July 2001 (pp. 126–135). Cambridge (UK): Springer.

- Hansch, C., & Leo, A. (1979). *Substituent constants for correlation analysis in chemistry and biology*. New York: Wiley.
- Kuhl, F. S., Crippen, G. M., & Friesen, D. K. (1983). A combinatorial algorithm for calculating ligand binding. *Journal of Computational Chemistry*, 5(1), 24–34.
- Moon, J. W., & Moser, L. (1965). On cliques in graphs. *Israel Journal of Mathematics*, 3(1), 23–28. doi:10.1007/BF02760024.
- Reymond, J.-L., & Awale, M. (2012). Exploring chemical space for drug discovery using the chemical universe database. *ACS Chemical Neuroscience*, 3(9), 649–657. doi:10.1021/cn3000422, PMID: 23019491.
- Raymod, J. W., Gardiner, E. J., & Willet, P. (2002). RASCAL: Calculation of graph similarity using maximum common edge subgraphs. *The Computer Journal*, 45(6), 631–644.
- Raymond, J. W., & Willett, P. (2002). Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *Journal of Computer-Aided Molecular Design*, 16(7), 521–533.
- Rhodes, N., Willett, P., Calvet, A., Dunbar, J. B., & Humblet, C. (2003). CLIP: Similarity searching of 3D databases using clique detection. *Journal of Chemical Information and Computer Science*, 43(2), 443–448.
- Stah, M., Mauser, H., & Hoffmann, F. (2005). Database clustering with a combination of fingerprint and maximum common substructure methods. *Journal of Chemical Information and Computer Science*, 45(3), 542–548.
- Tanimoto, T. (1958). An elementary mathematical theory of classification and prediction. *Internal IBM Technical Report*.
- Toropov, A. P., Toropov, A. A., Lombardo, A., Roncaglioni, A., Benfenati, E., & Gini, G. (2010). A new bioconcentration factor model based on SMILES and indices of presence of atoms. *European Journal of Medicinal Chemistry*, 45(9), 4399–4402.
- Ullmann, J. R. (1976). An algorithm for subgraph isomorphism. *Journal of the ACM*, 23(1), 31–42.
- Xu, J. (1996). GMA: A generic match algorithm for structural homomorphism, isomorphism, and maximal common substructure match and its applications. *Journal of Chemical Information and Computer Science*, 36(1), 25–34.
- Zhu, Y., Oin, L., & Yu, J. X. (2013). High efficiency and quality: Large graphs matching. *The International Journal on Very Large Data Bases*, 22(3), 345–368.

Generative Topographic Mapping Approach to Chemical Space Analysis

Dragos Horvath, Gilles Marcou and Alexandre Varnek

Abstract Generative Topographic Mapping (GTM) is a probabilistic, non-linear dimensionality reduction method, developed by C. Bishop et al. It essentially represents a fuzzy-logics-based enhancement of Kohonen Self-Organizing Maps (SOM). The probabilistic nature of this method is the source of the multivalent applications of GTM, which goes well beyond simple dimensionality reduction and visualization, but allows straightforward comparison of large compound libraries (in terms of diversity and coverage), supports regression or classification models with applicability domain control and herewith may serve as predictive tools of a large panel of properties (including polypharmacological profiles of bioactive compounds). A good predictive modeling implicitly validates a map and provides an objective criterion to select the best suited ones, out of the multitude of possibilities based on different initial molecular descriptors and user-defined mapping parameters. This multi-purpose “Swiss army knife” of dimensionality reduction may furthermore extract “privileged” structural patterns associated to bioactivities of interest, and hence contribute to an intuitive understanding of structure-activity relationships.

Keywords Generative Topographic Mapping · Dimensionality reduction · (Quantitative)Structure-Activity relationships (Q)SAR · Privileged patterns

D. Horvath (✉) · G. Marcou · A. Varnek
Laboratoire de Chimoinformatique, UMR 7140 CNRS – University of Strasbourg,
1, rue Blaise Pascal, 67000 Strasbourg, France
e-mail: d.horvath@unistra.fr; dhorvath@unistra.fr

G. Marcou
e-mail: g.marcou@unistra.fr

A. Varnek
e-mail: varnek@unistra.fr

1 Introduction

In data mining, items characterized by a large number of attributes can be conceived as points in a “data space” or “attribute space” (defined by the vector of its attributes) but cannot be directly visualized as such if the dimensionality D (the total number of attributes needed to characterize the instance) exceeds three. Or, complex instances such as molecules (or chemical reactions) in chemoinformatics typically require hundreds or thousands of specific attributes—henceforth called “molecular descriptors”, as customary in chemoinformatics—in order to conveniently capture the chemical information associated to a given compound. Understanding the neighborhood relationships between items—the analysis of the relative distances separating them in data space—is often of paramount importance to the understanding and exploitation of the knowledge provided by a set of example items, if item properties can be shown to comply with the neighborhood principle. This principle states that similar items (close to each other in data space) tend to display rather similar properties. In chemoinformatics, the “similarity principle” postulating that similar molecules will likely display similar (physicochemical and/or biological) properties (Johnson et al. 1988; Johnson and Maggiora 1990) is a key paradigm in chemistry, guiding the design and synthesis of novel analogues of properties close to the ones of state-of-the-art precursor compounds.

The similarity principle may be perfectly well exploited in arbitrarily high-dimensional descriptor spaces (Papadatos et al. 2009; Patterson et al. 1996), based on therein defined distance measures (metrics) quantitatively rendering the degree of dissimilarity (remoteness) of any two items. However, such approaches are frustratingly counterintuitive “black boxes”. The alternative—intuitive grasping of the neighborhood relationships from a 2D map of the initial space—requires a procedure to project the initial points onto a plane, in a way minimizing distortions of inter-item distance value. In the practice of chemoinformatics, projected inter-item distances need not quantitatively match the ones in the initial descriptor space: it is enough to ensure that (i) neighboring molecules in the descriptor space continue to show up as neighbors, and (ii) initially remote species do not artefactually become neighbors in the projection (the so-called “latent space”). The principle of a meaningful projection is illustrated in Fig. 1—item a is closest to items d , e and b , in both the initial space and the projection.

Many various dimensionality reduction algorithms do exist, starting from the classical linear algebra Principal Component Analysis (PCA) (Dunteman 1989), to various non-linear techniques such as Kohonen Self-Organizing Maps (SOM) (Kohonen 1984, 2001), Multidimensional Scaling (MDS) (Agrafiotis et al. 2001), Stochastic Embedding (Agrafiotis 2003), etc.

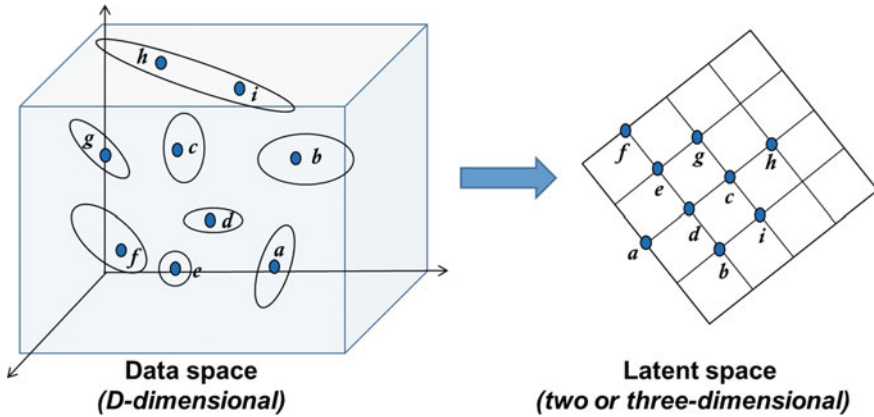


Fig. 1 The principle of dimensionality reduction

2 Generative Topographic Mapping—Principles

Generative topographic mapping (Kireeva et al. 2012) or GTM, introduced by Bishop et al. (1998a, b), are basically fuzzy-logics driven Kohonen Self-Organizing Maps (SOM). A regular squared grid of K nodes covering the 2D latent space is generated, where K is the square of some small integer \sqrt{K} , the grid “width”, expressed by the number of nodes/square edge. A node k is defined by its integer 2D coordinates $\mathbf{x}_k = (l_x, l_y)$, with index $k = l_x \times \sqrt{K} + l_y$ where $l_x, l_y = 0, \dots, \sqrt{K} - 1$ and $k = 0, \dots, K - 1$. Each node is mapped to a manifold point \mathbf{y}_k embedded in the D -dimensional space: $\mathbf{x}_k \rightarrow \mathbf{y}_k$, using the non-linear mapping function $y(\mathbf{x}; \mathbf{W})$ that maps points from the two-dimensional latent space into the D -dimensional data space:

$$y(\mathbf{x}; \mathbf{W}) = \mathbf{W}\varphi(\mathbf{x})$$

$$\mathbf{Y} = \mathbf{W}\Phi^T$$

where \mathbf{Y} is the $K \times D$ manifold, \mathbf{W} is the $D \times M$ parameter matrix, and Φ is the $M \times K$ radial basis function matrix with M RBF centers $\boldsymbol{\mu}_m$:

$$\Phi_{mk} = \exp\left(-\frac{\|\mathbf{x}_k - \boldsymbol{\mu}_m\|^2}{2\sigma^2}\right)$$

The parameter σ^2 corresponds to the average squared Euclidean distance between two RBF centers, multiplied by a tunable factor w . \mathbf{W} , the parameter matrix, can be initialized such as to minimize the sum-of-squares error between initial-space and latent-space point distances, corresponding to a default, linear PCA mapping. The points \mathbf{y}_k on the manifold are the centers of normal probability distributions (NPDs) of \mathbf{t} :

$$p(\mathbf{t}|\mathbf{x}_k, \mathbf{W}, \beta) = \frac{\beta^{D/2}}{2\pi} \exp\left(-\frac{\beta}{2} \|\mathbf{y}_k - \mathbf{t}\|^2\right)$$

where \mathbf{t}_n is a data instance and β the common inverse variance of these distributions.

Intuitively, one may think of this abstract manifold as a “rubber sheet” is inserted in the descriptor space zone covered by the items to map, and which may be subsequently “torn” (fitted) in order to accommodate, or at least pass closely to each of these items. Therefore, the ensemble of N data items—here, N molecules, each represented by their molecular descriptor vector $\mathbf{t}_n, n = 1, \dots, N$, define the zone within which the manifold will be most accurately defined, thus represent a “frame” within which the map is positioned and will therefore be termed “frame set” in the following. Manifold grid points are positioned in the neighborhood of frame set points. The optimization of the below-given log likelihood function accounts for the purposeful distortion of the manifold in order to optimally cover or approach each of the frame set items.

Eventually (see intuitive example in Fig. 2), the manifold will be folded into the delimited 2D square grid of K nodes. While it may describe an infinite hypersurface in the initial space, its extrapolation far beyond the frame set-covered zone is hardly meaningful. “Exotic” items outside the frame zone will be spuriously “folded” back within the bounds of the square grid, but should be eventually ignored, because they are outside the applicability domain of the map. Therefore, the proper choice of frame compounds—which may, but do not need to coincide with the actual compound collections targeted by the GTM-based study—is a key prerequisite in GTM design.

An optimal GTM corresponds to the highest log likelihood \mathcal{L} , taken over all frame compounds $n = 1, \dots, N$, optimized by expectation-maximization (EM):

$$\mathcal{L}(\mathbf{W}, \beta) = \sum_n \ln \left\{ \frac{1}{K} \sum_k p(\mathbf{t}_n | \mathbf{x}_k, \mathbf{W}, \beta) \right\}$$

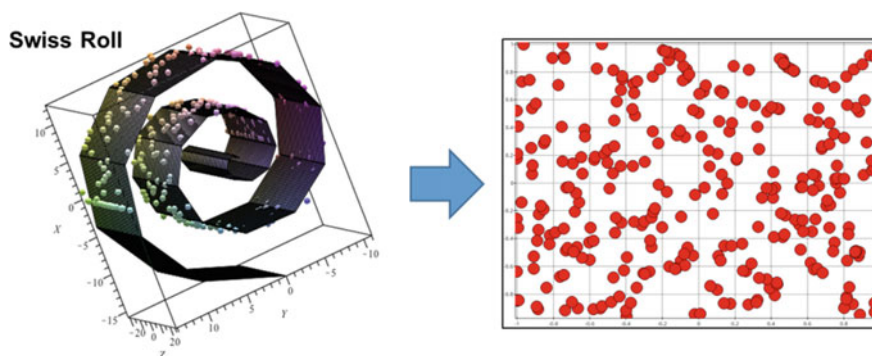


Fig. 2 An example of the “Swiss roll” manifold fitted to match the items (*points*) in the initial 3D data space, then unfolded onto the 2D latent space grid

β and \mathbf{W} are optimized during the maximization step:

$$\frac{1}{\beta} = \frac{1}{ND} \sum_n \sum_k R_{kn} \|\mathbf{y}_k - \mathbf{t}_n\|^2$$

$$\left(\Phi^T \mathbf{G} \Phi + \frac{\lambda}{\beta} \right) \mathbf{W}^T = \Phi^T \mathbf{R} \mathbf{T}$$

where \mathbf{I} is the identity matrix and \mathbf{G} a $K \times K$ matrix with elements $G_{kk} = \sum_n R_{kn}$.

GTM build-up is fully controlled by four user-defined parameters: M , the number of RBFs, the number of nodes K , the RBF width multiplication factor w and the weight regularization coefficient λ . The latter two serve to tune the stiffness of the manifold and hence avoid overfitting. Of course, the local minimum which will be reached by the (gradient-based) optimization of the log likelihood function will depend on the initial geometry of the manifold—in our implementation, optimization starts from a flat manifold representing the plane of the first two principal components of the descriptor space. Note that the final rendering of the map may depend a lot on the initial conditions—but the neighborhood relationships it encodes will *not*: compounds that are close in the descriptor space should be mapped to adjacent points in latent space. Or, in chemoinformatics GTMs serve to monitor neighborhood relationships—therefore, no systematic study of all possible log likelihood minima achievable for any given quartet of control parameters (M , K , w , λ) has been pursued.

Eventually, the responsibility or posterior probability that a point \mathbf{t}_n in the data space is generated from the k th node is computed using Bayes' theorem:

$$R_{kn} = p(\mathbf{x}_n | \mathbf{t}_k, \mathbf{W}, \beta) = \frac{p(\mathbf{t}_n | \mathbf{x}_k, \mathbf{W}, \beta) p(\mathbf{x}_k)}{\sum_{k'} p(\mathbf{t}_n | \mathbf{x}_{k'}, \mathbf{W}, \beta) p(\mathbf{x}_{k'})}$$

These responsibilities are used to compute the mean (real value) position \mathbf{xy} of a molecule on the map $\mathbf{x}(\mathbf{t}_n)$, by averaging over all nodes with responsibilities as weighting factors:

$$\mathbf{xy}(\mathbf{t}_n) = \sum_k \mathbf{x}_k R_{kn}$$

Each point on the GTM thus corresponds to the averaged position of one molecule. This step completes mapping, i.e., reducing the responsibility vector to a plain set of 2D coordinates \mathbf{xy} , defining the position of the projection point of the initial D -dimensional vector on the map plane.

The above optimization of \mathcal{L} requires access to the entire set of molecular descriptor vectors, which represents a $N \times D$ matrix of real numbers, and may thus quickly run into memory problems, knowing that the dimensionality D of the

descriptors may often reach order of magnitude of 10^3 – 10^4 , whilst frame sets of millions of molecules could be considered. In particular, in the context of “big data”, or in order to exhaustively map the chemical “Universe” of all commercially available or feasible compounds, the input data alone may easily scale up to tens of GB, and temporary variables required for processing will be similar in size. Alternatively to the use of a memory-rich supercomputer, an incremental version of GTM (iGTM) algorithm has been proposed (Gaspar et al. 2014). It divides the data into blocks and updates the model block by block until convergence of the log likelihood function.

After a map has been “built”—i.e., the manifold was optimized, based on provided frame set items—any other point \mathbf{t}' in the initial descriptor space can be projected on the manifold and its responsibility vector can be computed. This is technically possible—but practically not advisable—even if the compound is very different from frame set molecules, and implicitly remote from the fitted manifold. Note that mapping of external items only requires the manifold equation and \mathbf{t}' as input, thus can be easily parallelized, so that arbitrarily large external compound sets can be mapped on any given GTM. However, the underlying—smaller—frame set must be nevertheless representative of these external sets (i.e., cover roughly the same descriptor space hypervolume, albeit at much lower density). This is a key issue in chemical space mapping, needed to ensure that mapping of external compounds is meaningful, and not artefact-prone.

2.1 Responsibility Patterns

GTM has, over other techniques, the key advantage of its two-stage approach to dimensionality reduction:

1. from the original, D -dimensional descriptor space to the K -dimensional responsibility vector space (Responsibility Level). A responsibility vector (Fig. 3) can be intuitively visualized by colored “patches” positioned at the node (s) with significant responsibility values, where color intensity is modulated by the actual responsibility values.
2. from responsibilities to 2D positions on the map (2D Level): computed latent coordinates $\mathbf{xy}(\mathbf{t}_i)$ can be assigned to each compound: see crossfires at the (responsibility-weighted) barycenter of the set of significant residence nodes in Fig. 3.

The latter and final level is clearly not the most interesting one. A 2D map of very high-dimensional spaces will—irrespective of the linear or non-linear mapping strategy—be inherently imprecise, to the point of not being of great use. Molecular structures cannot be robustly characterized by two real numbers only, irrespective of the strategy one may design for defining those numbers. The full advantages of GTM are apparent at responsibility level, at intermediate

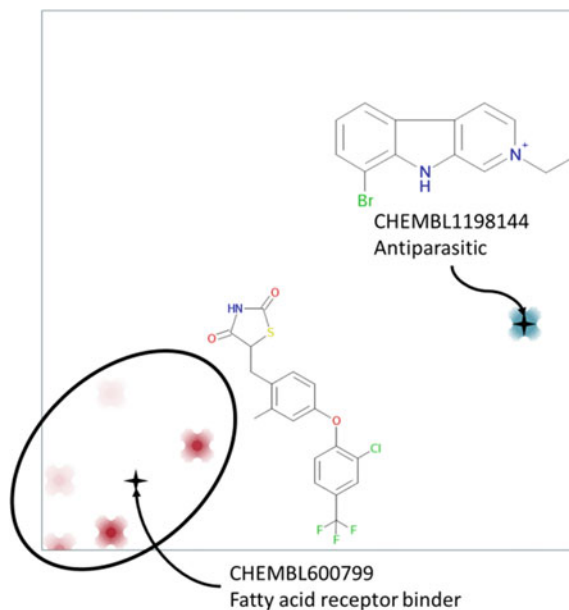


Fig. 3 Examples of single-node residents versus fuzzy, multi-node residents. For the antiparasitic compound above, the responsibility vector is null for all nodes except the highlighted one, in which the compound is predicted to reside exclusively. The fatty acid inhibitor below is defined as a partial resident of the highlighted nodes in *red*, where the color intensity matches relative responsibility values. The *black* crossfire signs correspond to the (x, y) latent space coordinates of the compounds of the map, and are positioned at the (responsibility-weighted) barycenter of the set of significant residence nodes. The displayed map is the result of a study (Sidorov et al. 2015) aimed at the discovery of general maps of maximal pertinence for the space of drug-like compounds (“universal” map #2 of the cited publication)

K dimensionality. K is a user-tunable parameter, which should be chosen such as to avoid massive loss of chemically relevant information, but filtering out the noise due to less relevant descriptor components.

It is straightforward to expect that similar molecules are to be represented by similar responsibility “color patches” on the map, and the human eye is perfectly suited to detect “color patch” similarity—even beyond the trivial scenario when “patches” include a single node and the GTM acts like a classical Kohonen map. Further reduction of the molecule object to a single point of 2D coordinates (crosshair), which is precisely the barycenter of the responsibility pattern, may represent a drastic loss of information, unless one single node accounts for the entire density distribution.

Fuzzy compound-to-node assignment may seem like a minor enhancement, but is actually another key strength of GTM over Kohonen maps. First, at a same grid size K , the volume of chemical information that can be monitored by a GTM is much larger. A Kohonen grid of K nodes may distinguish between at best K different core structural motifs—much less in practice. Some of these K nodes will,

indeed, each stand for “main stream” compound classes, but others will serve as “garbage collectors” of all the exotic structures that would not fit any of the former, but need to be assigned to one (and only one) given node, nevertheless. On a GTM, molecules are not necessarily bound to a single node and the total number of distinct structural motifs is defined—intuitively—by the number of distinct “color patch” patterns that may be drawn with the help of a K -sized grid, or—technically—by the phase space volume spanned by the K -dimensional responsibility vectors. Kohonen maps operate only with pure states, while GTMs, by contrast, with mixed states, and the latter come in virtually infinite numbers (not all of them corresponding to real compounds or common core motifs, however). A consequence is that exotic compounds that are remote from all the nodes of the manifold will as a consequence be often mapped with equally weak responsibilities on all nodes, rather than assigned to the one—relatively—closer “garbage” node.

However, the non-fuzzy Kohonen maps seem to have an apparent advantage in terms of compound clustering. All compounds mapping to a same node are, from the Kohonen map perspective, no longer distinguishable and therefore may be unambiguously viewed as members of a same group, or cluster—which makes perfect chemical sense for all but above-mentioned “garbage” nodes. Conceptually, things are identical for compounds residing in single nodes of GTMs, with the additional benefit that single-node residents are typically compounds found close to the manifold, well within the GTM applicability domain. Single-node residents of any given node are expected to form a chemically meaningful “cluster” of similar compounds. The cluster corresponding to the “blue” node in which resided the antiparasitic compound in Fig. 4 is shown below.

Yet, the working hypothesis “compounds of a same node belong to a same cluster” may be easily generalized to fuzzily mapped items such as ChEMBL600799 from Fig. 3, by introducing the concept of Responsibility Patterns, RP. The responsibility pattern (Klimenko et al. 2016) of a compound n is defined as an integer, discretized version of the real-number responsibility vector R :

$$RP_{kn} = [10 \times R_{kn} + 0.9]$$

where “[]” stands for the truncation operator. The peculiar transformation rule above was chosen such as to ensure that even marginally responsible nodes (at $R_{kn} = 0.01$) will be highlighted. Beyond this minimal threshold, every additional 10% increase of a responsibility value contributes an increment of +1 to the integer RP equivalent. For a compound n , the responsibility pattern vector RP_{kn} may be best rendered as *string* enumerating—in increasing node number order—the nodes of non-zero RP values, concatenated to these values, e.g., /k1:RP_{k1n}/k2:RP_{k2n}/k3:RP_{k3n}/.../. For single-node compounds, the RP string /k:10/ is simply a label of the concerned node. Herewith (see Fig. 5), compounds associated to a same RP string will be considered to belong to a same cluster.

The responsibility pattern approach therefore amounts to a cell-based clustering technique: the above discretization formula may be interpreted as a procedure to tessellate the vector space of responsibilities R , so that items within any cell would

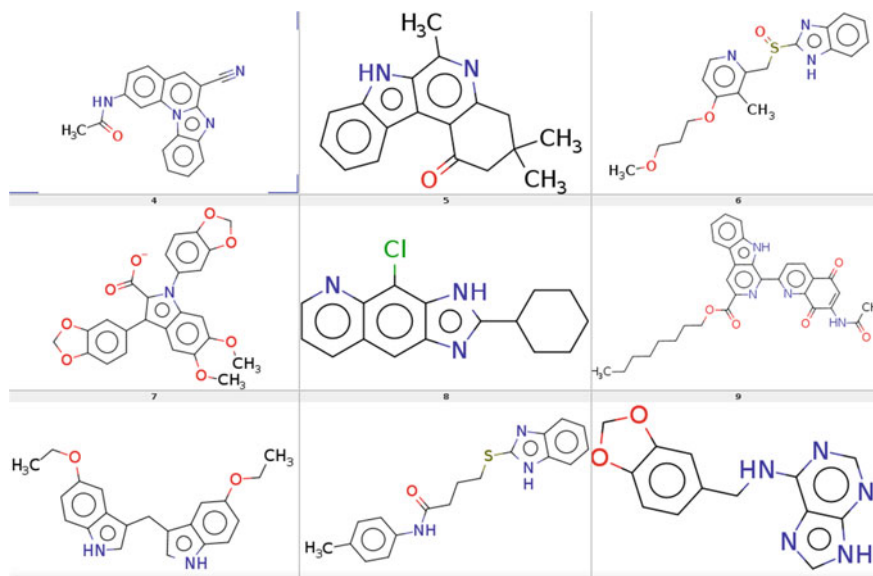


Fig. 4 Other single-node resident compounds from the ChEMBL database (Gaulton et al. 2011), in the node of residence of the antiparasitic compound from Fig. 3. The cluster regroups benzopyroles, benzimidazoles and other closely related heterocyclic scaffolds. In spite of the wide diversity of substituents (only 9 randomly picked examples are shown, out of its 9100 members in the ChEMBL database), there is a clearly visible common structural pattern associated to this node

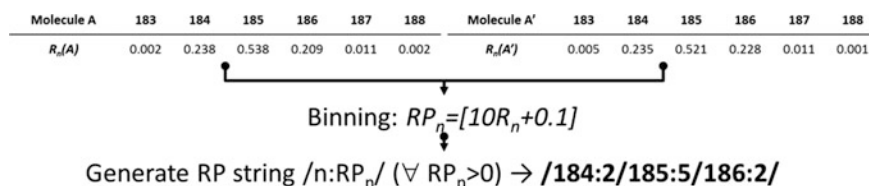


Fig. 5 Exemplifying the definition of Responsibility Patterns as strings (labels) concentrating the information in the responsibility vectors by “binning”, and herewith regrouping molecules with identical or slightly different responsibility vectors under a common label. Integers *above the lines* are node numbers, and corresponding real values below are responsibility values. These are binned, and nodes returning non-zero binned values are concatenated together with their binned value, into a “RP string” shown below

share a same *RP* string. Typically, cell-based clustering fails in high-dimensional spaces, because of the sheer number of possible cells: in the K -dimensional space of responsibility vectors, there are 10^K possible cells, with $K = 32 \times 32 = 1024$ in the GTM from Fig. 3 (see figure caption for more information about the map). However, with a well-fitted GTM model, only a minority of cells is actually populated—in particular, the K single-node configurations and fuzzy configurations with responsibilities shared between two and—for CHEMBL600799, which was

picked for being the “fuzziest” mapper amongst all ChEMBL compounds—five participating nodes. The 1.3M ChEMBL compounds populate 23,253 distinct responsibility patterns, out of which 723 correspond to strict single-node mapping modes (concerning a total of 1.22M compounds, i.e., 95.3% of ChEMBL structures). There are 19,952 bi-nodal *RPs*, regrouping some 53K molecules. There are 4,217 ChEMBL compounds that are nonspecifically “smeared” over the entire map, with $R_{kn} = 1/1024$ for all k —these 3% represent mapping failures, and are beyond the applicability domain of the model.

Like in any clustering approach, the user expects to see “structurally related” compounds grouped together under a given RP label. “Structural relatedness”, however, is an intrinsically ill-defined concept—it typically refers to the scaffold-centric view cherished by medicinal chemistry, where two compounds are “analogues” if they contain a same (however defined) “scaffold”. There is no absolute truth in the above point of view—one may as well prefer the alternative pharmacophore-centric approach, where two compounds are “analogues” if they are porters of a same pharmacophore pattern (analogous spatial distribution of functional groups of analogue physico-chemical nature). Note that the GTM-based RPs are not specifically generated on the basis of scaffold-centric information, but may capture the presence of a scaffold by its specific “signature” in the provided molecular descriptor vector (specific scaffold contributes a subset of specific fragments to the ISIDA fragment count vector). Alternatively, pharmacophore patterns might also be captured, if the pharmacophore-colored fragmentation schemes (Ruggiu et al. 2010) are enabled.

Thus, the nature of the common structural “motif” behind a given RP is by default open-ended. First, “garbage” RPs—the equivalent of Kohonen “garbage” nodes—may appear, for various reasons. They may regroup cases of exotic compounds that are too far from the manifold to be clearly assignable to a node and are therefore “smeared” over many putative locations. However, single node RPs may also sometimes accommodate a set of—from a chemist’s point of view—highly diverse structures, with no obvious common “motif”. The more populous a node, the higher are its chances to accommodate widely diverse compounds. Figure 6 highlights the three most populous RPs of the ChEMBL map, each corresponding to single node RPs associated to the pinpointed “borderline” nodes. In spite of the large compound populations, in two of the three nodes it was rather easy to evidence the common underlying structural pattern “uniting” these compounds into a cluster. Finding the pattern required nothing but visual inspection of some representatives. Then, the observed putative common hypothesis were formulated as substructure search queries, and applied to the compounds matching each RPs—as their numbers are too large for exhaustive visual inspection. Indeed, 95% of the members of the most populous RP of ChEMBL (highlighted node #128) are putative Michael acceptors, matching the α,β -unsaturated ketone pattern $C=C-C=O$. More than 66% of residents of node #32 are oxyanions—which is a remarkable enrichment, knowing that over the entire ChEMBL set, the occurrence rate of oxyanions is of 17%. However, there is no obvious common pattern within

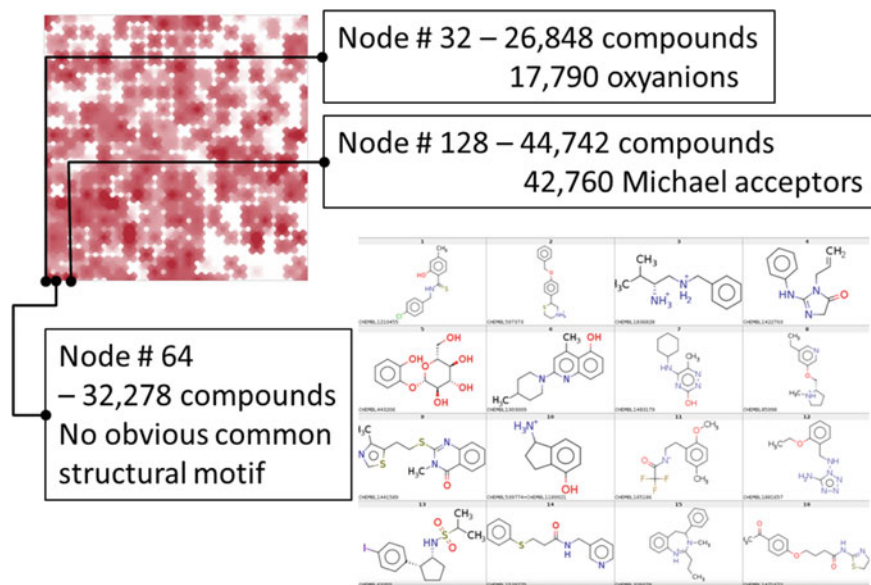


Fig. 6 Analysis of the three nodes corresponding to the three most populous Responsibility Patterns in ChEMBL, based on the map introduced in Fig. 3. Given the structural diversity of compounds in node #64, this may be viewed as a “garbage” node—nevertheless, it has the specificity of regrouping small, fragment-like compounds

node #64, the second-largest RP in ChEMBL. Yet, the molecules do have something in common—their “fragment-like” size, being significantly smaller, and hence less complex than typical drugs.

Thus, as exemplified in this chapter, and as observed in previous works (Klimenko et al. 2016), the unifying structural “reasons” behind a given RP may be of diverse nature, and represent different “resolution” levels. They may range from the extremely fuzzy size considerations, to clustering molecules by their predominant pharmacophore feature—anionic nature, to specific shared substructures.

These substructures *need* not to be scaffolds in order to be (bio)chemically relevant. As shown above, the herein discussed GTM “spontaneously decided” to regroup Michael acceptors, based on the specific signature of the $C=C-C=O$ moiety. Of course, the mapping process did not rely on any knowledge of putative specific or non-specific biological effects, “PAINS” (Bael and Walters 2014; Dahlin et al. 2015) of Michael acceptors. Also, it cannot be taken as granted that Michael acceptors would, as such, share a specific zone in the initial descriptor space—which is a prerequisite for their projection onto a common RP. The nature of molecular descriptors on which mapping was based (Sidorov et al. 2015)—force-field-type colored ISIDA atom pair counts—was of paramount importance, because they helped to evidence the specific signature of the $C=C-C=O$ moiety. This map was grown and selected with respect to its propensity to explain

structure-activity relationships throughout diverse series of compounds associated to various targets (vide infra—Building high-quality GTMs). In the process of achieving that goal, assignment of putatively reactive and hence unspecific Michael acceptors to a “borderline” node emerged spontaneously. Note—not shown in Fig. 6—that the fifth-most populous node of the map is another Michael-acceptor dominated chemical space zone, with the peculiarity that the $C=C-C=O$ pattern is now included in a ring.

Common structural motifs may nevertheless correspond to one scaffold, or to ensembles of similar scaffolds, as already highlighted in Fig. 4. Examples from previous work show that the relevant underlying common substructure may be more stringent than the scaffold level—compounds within a given RP may share not only a common scaffold, but also very specific substituents at key scaffold positions. The relationship between RPs and the underlying structural motifs is therefore open-ended and self-adaptive: it may stretch from very fuzzy regrouping of compounds sharing a same small size, or a same negative charge, to compound clusters based on a clear-defined common substructure, which may or may not match a scaffold (in the sense of “ring system”). Different maps may highlight different structural motifs that are specific to some of their RPs. On the contrary, “rediscovery” of the very same ones being associated to map-specific RPs (Klimenko et al. 2016) of different maps is also possible, even if those maps are based on different molecular descriptors. Either way, if a map is shown to be neighborhood-compliant, in the sense of supporting robust structure-activity models for a wide panel of properties (vide infra), then the RPs extracted from such map are highly likely to correspond to some well-defined underlying structural motif of (bio)chemical significance.

2.2 *GTM-Driven Classification and Regression Predictive Models*

Whenever molecules, initially represented as D -dimensional objects in descriptor space, are mapped onto a 2D latent space, their properties are being implicitly localized on the map. It makes sense to “transfer” the—mean—property of molecules residing a given latent space zone to that particular latent space zone itself. If the mapping is meaningful—that is (Horvath and Barbosa 2004), similarity principle-compliant—for a given property, then mappers onto any given latent space “spot” of sufficiently small size will be similar molecules of similar property. Hence, the mean \bar{P} of these property values will display a limited standard deviation $\sigma(P)$, and coherently represent the local, above-expectation accumulation of compounds of property value $P \approx \bar{P}$. Coherence of mapping of compounds with known property values may thus serve to implicitly define the quality of a mapping approach. Moreover, would a new species be shown to map in the same latent space zone, the assumption that its expected property value shall not be far from \bar{P} can be upheld and used for prediction.

P may stand for various properties of distinct nature—both continuous and categorical (class labels). The relevant latent space “spots” and the “mean” values \bar{P} may be defined in context-dependent ways, but the above outlines the general principle of predictive mapping. For example, in Kohonen maps, nodes are the smallest addressable latent space unit for which \bar{P} values may be computed. A Kohonen map may not make any more detailed prediction but returning the \bar{P} value associated to the node into which a compound has been classified. This means that, with continuous properties, it may return a discrete spectrum of node-bound \bar{P} values, one value per node for all “non-garbage” nodes, with $\sigma(P)$ below some user-defined threshold. Therefore, Kohonen maps—unlike GTM—fail to support proper quantitative regression models: they would return \bar{P} as the predicted value for the entire series of analogues residing in the same node. A structural modification of a compound would *not* trigger any change of the predicted value \bar{P} unless this change causes relocation to a different node, associated to a different \bar{P} value. The above, however, is perfectly compatible with the expected behavior of a classification model. Both Kohonen and GTM approaches may therefore be used for compound classification, while GTM is—due to its fuzzy mapping abilities—better suited for regression models.

When defining the “mean” property value \bar{P}_k of a GTM node k , one must count each resident compound n proportionally to its degree of residence in that node, R_{kn} :

$$\bar{P}_k = \frac{\sum_n R_{kn} \times w(n) \times P(n)}{\sum_n R_{kn} \times w(n)}$$

where $P(n)$ represents the property of compound n and $w(n)$ represent importance weighting factors of the compounds. When the property P represents a continuous magnitude, such as a pIC_{50} or $\log P$ value, there is no immediate reason for any specific importance weighing scenario. Letting all compounds be equally important, $w(n) = 1 \forall n$, will assign simple arithmetic means of the property to nodes. Since R_{kn} is never strictly zero, no matter how far compound n is situated from manifold node k in descriptor space, GTMs—unlike Kohonen maps—do not display genuinely empty nodes, and the above equation is applicable for all k , without fearing divisions by zero. However, if the above denominator is low, it makes little sense to expect a meaningful extrapolation of \bar{P}_k based only on the remote contributions of compounds having no significant degree of residence at k . Therefore, there should be some user-defined minimal threshold for the total cumulated responsibility per node, below which k should be considered as “practically empty”, and its technically obtainable but chemically senseless \bar{P}_k value ignored. Node density can be encoded in plots by color transparency—from completely transparent (below defined density threshold) to full color (to be used, for example, for the top $t\%$ most dense nodes). Density (cumulated responsibility) is a major criterion of the *trustworthiness* of estimated \bar{P}_k values: the higher the density, the more robust the assigned \bar{P}_k and the above-mentioned minimal density threshold can be considered

as an applicability domain delimiter of a map (Gaspar et al. 2015, 2013; Horvath et al. 2009; Sushko et al. 2010; Tetko et al. 2008).

Note that the equation above may also be used for classification purposes: if, for example, we define $P(n) = 1$ for all inactives, by contrast to $P(n) = 2$ for all actives, then \bar{P}_k above will be below 1.5 if the inactives residing the node are predominant, above 1.5 if actives are dominant and 1.5 if both categories are equally well represented. Then, the node can be assigned to either class 1 or 2, or discarded as “undecidable”. Note that the herein—for the sake of intuitiveness—outlined approach to classification by rounding up the \bar{P}_k values only works properly for two-class classification problems. Obviously, with three classes $P(n) \in \{1, 2, 3\}$ a node having 50% of its residents of class 1, and the other half of class 3 would be wrongly colored as “class 2”. The proper way (Gaspar et al. 2013) to deal with multi-class classification is to count cumulated responsibilities $\sum_n R_{kn} \times w(n)|_{n \text{ of class } P}$ for each class, and to return the class P with the largest sum as \bar{P}_k . Only two-class classification supports node coloring by the fuzzy \bar{P}_k gradually shifting from class 1 to class 2, and herewith implicitly returning the *coherence-based trustworthiness* of node versus class association—the closer \bar{P}_k is to the extremes 1 or 2, the more robust the prediction. Coherence-based trustworthiness can also be defined for multi-class classification problems, by checking how much larger the winning cumulated responsibility score is with respect to the second-best one. Coloring nodes by winning class only is always feasible, but unfortunately such plot does not inform about coherence-based trustworthiness. In regression models, coherence-based trustworthiness can be inferred from the standard deviation $\sigma_k(P)$ of the property at node k (vide infra). Two-class classification is a special case, since $\sigma_k(P)$ is deterministically related to \bar{P}_k : it is zero if \bar{P}_k reaches its extreme values 1 or 2, and maximal (equal to 1) when \bar{P}_k is an undecided 1.5. Fuzzy class landscapes rendering the mean class \bar{P}_k of GTM nodes have the peculiarity of informing both about the winning class in each node *and* the coherence-based trustworthiness of that assumption. With transparency encoding local compound density, they provide a complete picture of the class landscape and its applicability domain.

2.2.1 Bayesian Weighing to Correct for Class Size Imbalance

If the classes to be discriminated against on a map are of very different sizes (classically, the number of inactives in screening of random compound collection is much higher than the one of confirmed actives, for example) then it may be interesting to revisit the discussion above in the light of *relative*, rather than absolute predominance of a class in a node. Practically, if the default ratio of actives versus inactives is 1:100 throughout the studied compound collection, then a node populated by one active for only five inactives is still dominated by inactives in terms of absolute “head counts”, and yet *enriched* in actives by a robust factor of 20. Therefore, it may deserve to be highlighted as a node of “actives”, nevertheless.

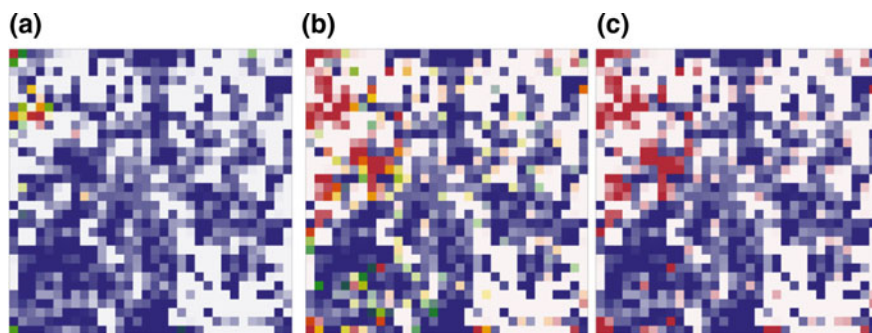


Fig. 7 GTM node coloring (each of the 36×36 nodes being a small squared “tile” of the grid) by average fuzzy class value (plots *a* and *b*), where the five-color spectrum maps P_k values from 1 (aliphatic class) in *red* to 2 (aromatic) in *blue*, with *middle* color *yellow* marking “undecidable” nodes. Plot *a* is realized in terms of absolute compound numbers per class (out of the 1.3M monitored ChEMBL compounds), whilst plot *b* monitors the relative enrichment of nodes in terms of aliphatic and aromatic compounds, respectively. Plot *c* represents the simplified two-color “winning class” landscape of *b*, where the “undecidable” nodes turn either *red* or *blue*, depending on whether their P_k value was slightly larger or smaller than 1.5. Node transparency is modulated by their cumulated responsibilities, i.e., the fuzzy count of resident compounds

This can be achieved by corrected importance weights for molecules of each class, taking their default occurrence rates as baseline. With i classes, denoting the total fraction of class i members of the library by f_i , the weight $w(n)$ of a compound n belonging to class i should be set to $w(n)|_{n \in i} = f_i^{-1} / \sum_j f_j^{-1}$. In this way, nodes would be “undecidable” at $\bar{P}_k = 1.5$ if their relative population of the two classes equals default occurrences, “active” if the population of actives is higher than the default “hit rate” or inactive otherwise. Figure 7 illustrates this aspect in monitoring the distribution of ChEMBL aromatic versus aliphatic compounds on the GTM nodes (same map as in Fig. 3, see references in that figure legend). Out of the 1.3M ChEMBL compounds, only 83K are completely void of aromatic moieties and were labeled “aliphatic” (class 1), whilst the vast majority of aromatic moiety-containing molecules are considered in class 2. Plots *a* and *b* below correspond to the plain $w(n) = 1$ scenario and the occurrence-based importance weighing, respectively. The five-color spectrum maps \bar{P}_k values from 1 (aliphatic class) in red to 2 (aromatic) in blue, with middle color yellow marking “undecidable” nodes. It can be seen from plot *a* that nodes in which the purely aliphatic compounds significantly outnumber the ubiquitous aromatic derivatives are rare. If occurrence rate-based importance weights are used, nodes *relatively* enriched in aliphatics are witnessing a “red shift” of their colors. Eventually, plot *c* of the same figure is identical to plot *b*, but in bicolor mode highlighting only the winning class color.

Plot *b* is clearly the most informative of the three alternative renderings—it shows that aromaticity/aliphatic character, a physicochemical parameter of key relevance in drug design, defines a major “fault line” on the map, with aliphatics relatively predominant in the north-west. It also gives a clear demarcation of nodes

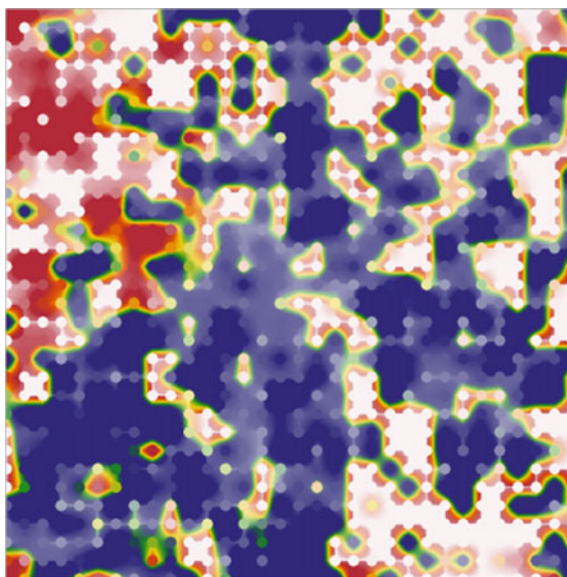
that are robustly dominated by one or the other class, versus mixed ones—which will nevertheless be forcibly declared “aliphatic” or “aromatic” in the traditional “winning class” representation c .

Now, the plots in Fig. 7 might have represented a Kohonen map as well as a GTM, since they do focus only on the information associated to the nodes. At this level, the fact that compounds may be fuzzily shared by several nodes would not significantly impact the generic aspect of such plots. On a Kohonen map, a compound is assigned to a node, so it does make sense to show nodes as tiles covering the map. On a GTM, however, a compound shared between several nodes may be imagined as “residing”—in terms of (x, y) latent space coordinates—*between* the nodes, as shown in Fig. 3. Therefore, logically, the mapped property landscape is also defined over the entire latent space between the nodes and may, in principle, written as a function $\bar{P}(x, y)$, to be interpolated—according to various strategies, from node \bar{P}_k values. For example, $\bar{P}(x, y) = \bar{P}_k |_{k=\text{nearest node to } (x, y)}$ is called the “local” extrapolation strategy. By contrast, in the “global” strategy $\bar{P}(x, y)$ associated to a compound n located at (x, y) is not directly inferred from latent space coordinates, but falls back to the responsibilities relying that peculiar resident to the nodes of given \bar{P}_k :

$$\bar{P}(n) = \sum_k R_{kn} \bar{P}_k$$

Above, the predicted property is now a smooth function of responsibilities, so the GTM-specific global property prediction strategy is a genuine regression method for prediction of continuous molecular properties. In Fig. 8, the landscape

Fig. 8 Interpolated aromaticity class landscape—*red* aliphatic, *blue* aromatic. Compare to its “node-only” rendering in Fig. 7b



$\bar{P}(x, y)$ is obtained by polynomial interpolation with respect to the values of the four surrounding nodes. In such a landscape, nodes would merely correspond to individual grid points, but, in order to highlight their special status, small circles of homogeneous color corresponding to the actual \bar{P}_k values are “cut” out of the smooth, interpolated landscape. Compare the interpolated, GTM-specific and fuzzy aromaticity/aliphaticity class *landscape* below to its “node-only”, Kohonen-map like counterpart in Fig. 7b.

2.2.2 Density, Coherence, Applicability

In order to conclude on the key issue of trustworthiness/applicability of GTM-based property landscapes, it is interesting to emphasize the standard deviation $\sigma_k(P)$ and respectively mean node-based property \bar{P}_k values are not correlated (except for two-class problems). However, the former—“coherence”, a strong indicator of the trustworthiness of \bar{P}_k values—may be alternatively used, for example, as the color transparency modulation parameter on the map, to produce alternative coherence/property landscapes, which may significantly differ from above-introduced density/property plots, and herewith provide an independent point of view to chemical space analysis. This is exemplified in Fig. 9, representing three different viewpoints to the octanol-water partition coefficient $\log P$ map of the 1.3M ChEMBL compounds. As there are no experimental $\log P$ values for the entire ChEMBL, calculated values provided by the ChemAxon tool *generateMD* (ChemAxon 2007) were used instead. A common property coloring spectrum is used: red for extreme hydrophilic $\log P \leq 0.0$, blue for extreme hydrophobes at $\log P > 6.0$, orange–yellow–green for the intermediate ranges. Plot *a* in Fig. 9 is the “classical” density-modulated representation, which conveys a first image of density-conditioned trustworthiness: empty zones (cumulating the equivalent of less than 1 compound/node, in terms of total responsibility) are obviously not able to

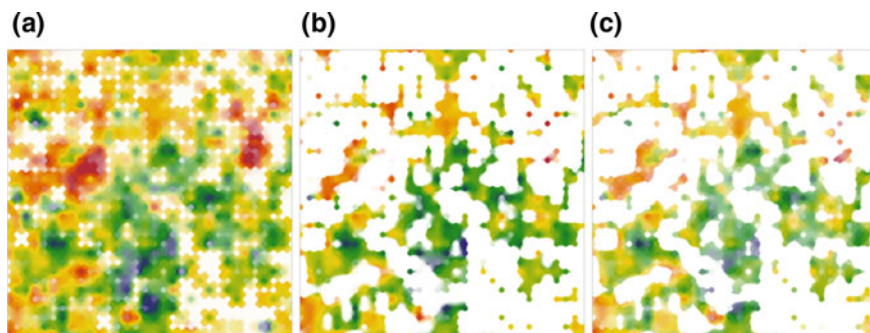


Fig. 9 Three alternative modes to represent the $\log P$ landscape of ChEMBL compounds: **a** density-modulated, **b** coherence-modulated, **c** applicability score-modulated. The used map has been introduced in Fig. 3

predict the lipophilicity of any external compound that might be mapped therein. By contrast, plot *b* is coherence-modulated: all nodes in which the standard deviation $\sigma_k(P)$ exceeds $2.5 \log P$ units are no longer visible, while those with $\sigma_k(P) < 1.0$ are fully colored. In general, low-density zones are also low-coherence zones. Therein, \bar{P}_k and $\sigma_k(P)$ are estimated on hand of remotely responsible compounds, that are basically “random picks” happening to be the less remote, not really descriptive of those zones, and therefore not expected to be coherent in terms of their $\log P$ values. However, there are significantly populated map regions that are *not* very selective in regrouping compounds according to their lipophilicity. Let us note, at this point, that the considered manifold was never built or selected (Sidorov et al. 2015) in order to maximize its predictive propensity of $\log P$. This notwithstanding, the map nevertheless features many zones in which compounds of roughly similar lipophilicity cluster “spontaneously”. Eventually, plot *c* below shows how density and coherence can be combined into a composite “applicability” parameter, defined as the product of density and a coherence penalty factor, reaching its maximum of 1 at $\sigma < 1.0$ and its minimum 0 at $\sigma > 2.5$. This applicability score, basically a coherence-modulated density, was used in plot *c* instead of “pure” density in plot *a*, all other setups being equal.

2.2.3 Building High-Quality GTMs—Properly Choosing Key GTM Parameters

Let us re-emphasize, at this point, that obtaining of property landscapes like above-shown is a process involving two clearly distinct steps:

1. the actual unsupervised map (manifold) construction, based on a frame set, and
2. subsequent (supervised) learning or “coloring” of this map, based on a—potentially different—training set.

Note, furthermore, that any manifold from step 1 may be, in principle, independently used in many alternative coloring attempts, in as far as the herein used training sets are not too remote from the frame-set-based manifold, as already mentioned.

Some options/parameters only concern only the unsupervised manifold fitting step 1. These include the four GTM setup parameters—node number K (required to be a perfect square integer, the number of radial basis functions (RBFs) M , RBF width factor w and weight regularization coefficient λ —in addition to the frame set and descriptor choices, which can be formally regarded as additional degrees of freedom, “meta-parameters”. By contrast, the choice of possible coloring/interpolation procedures required to build the property map does not affect at all step 1—any given manifold is in principle exploitable for both regression and classification, based on either above-mentioned “local” or “global” approaches.

All these (meta-)parameters have an impact on the quality of the final predictive model supported by the manifold. Model quality is a key objective criterion to

validate the quality of the proposed manifold. Without it, the “beauty” of a map is the only criterion to decide whether the chosen grid size is “correct”, whether the choice of a different set of molecular descriptors would have improved the mapping, etc.

Coupling visualization with prediction is therefore a key benefit of the GTM approach. Thus, one may formulate the GTM construction problem as a combinatorial optimization approach. Given all the possible choices of the seven already-mentioned (meta-)parameters (designation of a frame set and of a molecular descriptor type, out of the respective lists of possible choices, selection of the K , M , w and λ values and of the landscape interpolation strategy), which choice shall produce a map optimally rendering the one or more targeted property landscape(s)? It is understood that “optimal rendering” of a property landscape means maximizing the predictive power of such landscape. Placing an external compound (not used in the “coloring” process) on the colored map, in order to “read” the predicted property at the given location, is expected to return values in good quantitative agreement with experiment. Thus, map quality will be measured in terms of classical statistical validation criteria—cross-validated determination coefficients Q^2 , for example. To design a multicompetent map able to support more than a single predictive model, the “compromise” mean Q^2 might be used as a global criterion (optionally including a penalty for high standard deviations of Q^2 , in order to discourage setups with either extremely good or extremely bad results for the different monitored properties). One may alternatively consider a multiobjective optimization strategy, defining a Pareto front of locally best solutions for each of the monitored properties. The search for (near)-optimal setups in the seven-dimensional parameter space cannot be done systematically, knowing that the calculation of map goodness criteria may be a very time-consuming undertaking. Recall that this implies (1) fitting the manifold, given the descriptor choice, the frame set choice and the four GTM parameters, (2) cross-validated manifold coloring/prediction cycles, for each of the targeted properties, based on the property-specific training sets. Therefore, any stochastic search strategies—computer cluster-deployed genetic algorithms, for example—are well suited for optimal mapping parameterization.

Since a manifold needs not be tailor-made to specifically serve as support of a single dedicated model, one may ask whether it is possible to build some to successfully serve as support not only for the propertie(s) for which it was optimized (vide supra), but also for many other distinct and diverse structure-property models. So-far obtained results (Sidorov et al. 2015) of this quest for an arguably “Universal” GTM are very encouraging, having led to manifolds that showed to be valuable supports for hundreds of distinct predictive models, for properties as diverse and unrelated as target-specific activities, antiviral and antibiotic properties, physico-chemical properties. The maps used to exemplify the various issues discussed here are all, unless otherwise stated, “Universal” maps centered on the drug-like chemical space as represented in the ChEMBL database.

3 Chemical Space Analysis Using GTMs

The following will focus on the various ways of using GTMs for the rational and intuitive understanding of chemical space, and, implicitly, for library design. This covers topics as diverse as comparing different large compound libraries, or designing libraries with any desired coverage “pattern” of chemical space—both maximal diversity subsets, and focused libraries, putatively enriched in bioactives of desired class.

3.1 GTM-Based Compound Library Comparison

This topic has been extensively covered in previous publications (Gaspar et al. 2013, 2014, 2015), and therefore only a brief reminder of the underlying principles will be given here. The key concept here is representation of any compound library by the cumulated responsibility vector—the “density”—at any node. This renders any library, of arbitrary size, as a single K -dimensional vector, which is a mathematical object of the same class as (R_{kn}) , the molecular responsibility vector, i.e., the density vector of a “library” composed of one molecule, n . For a library L , the descriptor vector of cumulated responsibilities can be formulated as $(\sum_{n \in L} R_{kn})$. Therefore, two libraries, L and Λ , can be straightforwardly compared by means of taking some distance/dissimilarity score (Euclid, 1-Tanimoto, etc.) of their characteristic vectors, $(\sum_{n \in L} R_{kn})$ versus $(\sum_{n \in \Lambda} R_{kn})$. This is, first of all, extremely fast compared to calculating the pairwise inter-molecular dissimilarity scores of all members of L versus all members of Λ . If the distance metric is based on some covariance score which is independent of the absolute magnitudes of the two vectors, such as the cosine metric $\vec{x}\vec{y}/(\|\vec{x}\|\|\vec{y}\|)$, then two libraries with identical *pro-rata* representations in all GTM nodes will be reported as identical, irrespective of their sizes—as, for example, a representative “core” subset of a large collection versus this parent library. Library comparison can be intentionally rendered size-insensitive, all metrics confounded, by explicit normalization of cumulated responsibility vectors with respect to library size. If, furthermore, nodes were assigned mean characteristic property values \bar{P}_k , these may be used as weighing factors in library comparison metrics. In order, for example, to bias the library comparison with respect to the nodes which are enriched in actives for a given target— \bar{P}_k representing, for example, the mean pIC_{50} value of actives residing on node k —library comparison should use vectors $(\sum_{n \in L} \bar{P}_k R_{kn})$ versus $(\sum_{n \in \Lambda} \bar{P}_k R_{kn})$. This would implicitly focus more on the relative populations on nodes with high mean pIC_{50} values. Note that map “coloring” to obtain \bar{P}_k values need not be based on any experimental pIC_{50} of compounds from the actually compared L and Λ —any other independent “color” training set can be used to this purpose. Again, it is necessary to ensure, like always, that the used manifold is

“competent” to acquire the compounds of L , Λ , and the putative color set, as already discussed in the previous chapter.

Alternatively, library comparison can be formally treated like a classification problem. If compounds of L are, arbitrarily, considered of class 1, whilst Λ members are assigned the class label 2, then the fuzzy mean class \bar{P}_k associated to nodes will intrinsically reflect the (absolute or relative, *vide supra*) local dominance of either of libraries, in terms of density. A GTM image consisting of perfectly separated “patches” of red and blue means that the chemical spaces covered by the libraries does not overlap at all. A homogeneously yellow landscape, corresponding to $\bar{P}_k = 1.5 \forall k$ means that local densities of both libraries are quasi-identical all over the space. The former scenario would correspond to a Tanimoto score of 0, whilst the latter means Tanimoto = 1 in terms of cumulated responsibility vectors, as discussed above. In practice, one expects both zones of significant overlap *and* zones of separation to coexist: this would correspond to some intermediate score in terms of quantitative library comparison. However, the class landscape is much more information-rich than a simple Tanimoto score, because it conveys node-by-node information, rather than the final “verdict” condensed into a single score value. The left side of Fig. 10 represents such a class landscape, comparing the 1.3M ChEMBL compounds to a roughly equally large collection of 1.4M commercial compounds of various sources, curated for High-Throughput Screening compliance (Horvath et al. 2014). The “blue” chemical space zones that are clearly overpopulated with commercial compounds are well visible. Furthermore, comparing this class landscape to the lipophilicity landscape on the right side (same as

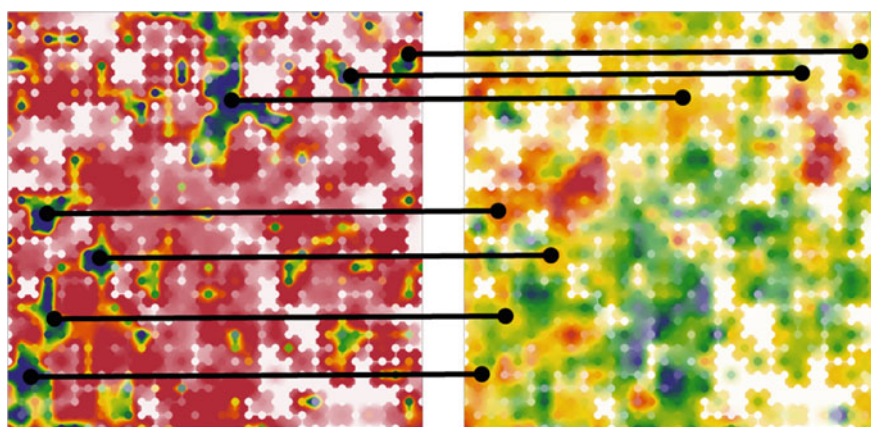


Fig. 10 (Left) Fuzzy mean class landscape (with Bayesian weighing) of the comparative map of 1.3M ChEMBL compounds (class 1, red) versus 1.4M curated molecules from commercial sources (class 2, blue). The used map is the one introduced in Fig. 3. (Right) The lipophilicity landscape already shown in Fig. 9a has been added aside for comparative purposes: it can be seen that the chemical space dominated by commercial compounds corresponds to several zones of moderated lipophilicity

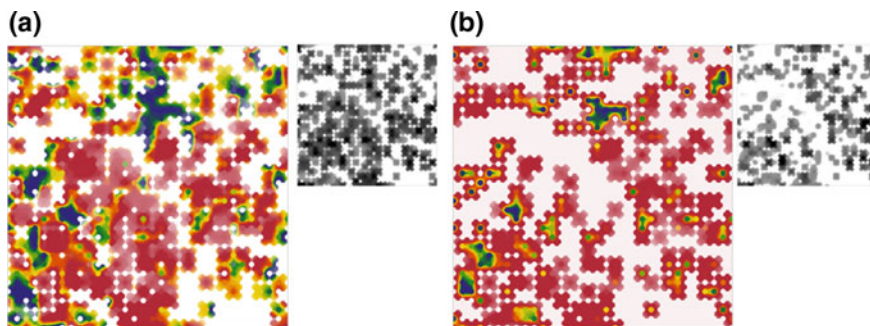


Fig. 11 Landscapes **a** and **b** represent the same “ChEMBL versus commercial” class landscape as in Fig. 10, now restricted to the compounds matching only Responsibility Patterns that were encountered, for at least ten times, amongst **a** the GPCR Sarfari and **b** the Kinase Sarfari ChEMBL subsets. Associated to **a** and **b** are the density plots of the cited ChEMBL subsets, in density-modulated *white–grey–black*

in Fig. 9a) immediately reveals that “commercial” chemical space is almost always associated to moderately hydrophilic compounds. It is, of course, straightforward to visualize representatives of either “blue” or “red” zones, as examples of collection-specific molecules.

Library comparison may furthermore be easily modulated and made to focus on peculiar chemical space zones. For example, in Fig. 11, the comparison of ChEMBL to the above-mentioned commercial compound library has been revisited from the perspective of two different medicinal chemists—one interested in GPCR research, the second active in the field of kinase inhibition. To this purpose, compounds of interest were selected for the given research domain—here, the predefined ChEMBL “SARfari” subsets for GPCRs and kinases, respectively. These were mapped, generating the white–grey–black density-modulated landscapes shown as miniatures below. The latter can be understood as problem-specific “masks” one would like to use in order to focus of chemical space zones of interest. Logically, this is the same thing as deciding to redefine the Applicability Domain of the map by means of the specific density of the “compounds of interest”.

Practically, the most straightforward way to apply such a filter is to

- extract the Responsibility Patterns (RPs) for all SARfari “compounds of interest”.
- establish a list of robustly reoccurring RPs, each representing at minimum 10 compounds of interest. On the herein used map, the ~115K compounds of the GPCR SARfari set cover 458 distinct RPs, while the less numerous (~51K) kinase SARfari compounds are responsible for 296 RPs.

- discard, from both ChEMBL and commercial libraries, all the compounds having RPs other than the ones kept in the above list.
- rebuild the fuzzy, mean class landscapes with the remaining representatives of the two libraries.

In the above-shown, compounds of interest were chosen to be rather large and heterogeneous sets, which are clearly not containing only actives with respect to the cited target class. However, focus on a wanted chemical space zone is extremely flexible: any set of RPs can be used, whether they come from validated bioactives of minimal potency, from compounds predicted to be actives by QSAR models, from promiscuous/specific compounds, etc.

3.2 GTM-Based Diversity Analysis

Let us consider the classical task of extracting a core subset of $c\%$ from a large library (here, ChEMBL) of maximal representativeness/diversity. GTMs are—like Kohonen maps—extremely useful for both proposing such a core subset, and a posteriori analysis of its relation to the (unselected remainder) of the parent library. Mapping in diversity analysis is a key time-saving step, because it provides an implicit “clustering” of molecules, by binding them to specific positions on the map. Molecules mapping to distinct locations—associated to different neurons on a Kohonen map, and, respectively, to distinct Responsibility Patterns (RPs)—are implicitly considered “diverse”. On the opposite, molecules which are assigned to a common location are indistinguishable, as far as mapping can tell. Here, GTM has the advantage of higher resolution: at equal number of nodes, the GTM supports more distinct RPs than the Kohonen approach, with their binary compound-to-node assignment scheme. A rational core extraction strategy supported by GTM would therefore amount to pick controlled numbers of compounds from the clusters associated to the detected RPs. This is extremely fast—the estimation of $O(N^2)$ intermolecular dissimilarity scores is completely avoided.

The most straightforward diversity selection strategy would therefore be a *pro rata* draw: in order to pick a representative core of $c\%$ molecules, it is advised to (randomly) pick $c\%$ of representatives of every detected RP. First, representatives of a given RP are, as already discussed, basically expected to be rather similar, and/or share some common structural traits. Therefore, in a “generic” library subsetting exercise, when there are no specified targets for screening the core library, there is little rationale to prefer one particular compound over all the other representatives of a given RP. Note that, in principle, one may use a classical diversity algorithm in the initial descriptor space (Agrafiotis 1997; Maldonado et al. 2006; Turner et al. 1997) for selection, ensuring that the RP-specific subset of $c\%$ avoids, as much as possible, inclusion of “redundant” compounds such as methyl/normethyl analogues. Even so, computer effort would remain reasonably low, since local comparison would

concern a limited number of items associated to a common RP. This was not pursued in this example, for three main reasons. First, the similarity threshold (Horvath et al. 2013) at which two similar molecules may be safely considered redundant is ill-defined and, at best, problem-specific. Second, the manner in which RP representatives are picked has no impact on the chemical space coverage as perceived by the GTM. Third, note that in practice two similar molecules may nevertheless happen to be assigned to different RPs because of binning artefacts. Map-based diversity selections are coverage-oriented, but do not formally guarantee the absence of redundant compounds. Therefore, if non-redundancy (whatever its definition) is a key issue, the optimal strategy is to generate a slightly larger-than-needed map-driven core selection, to be further refined by elimination of redundant compounds. This latter step will be relatively fast—since limited to the small core instead of the large library. Note that design of larger-than-needed cores is rather state-of-the-art protocol than exception. In practice, logistic bottlenecks, compound purity/solubility etc. will be highly impacting factors on the final compound selection. Therefore, diversity selection should be kept conceptually simple, and fast—extensive number-crunching coming up with an ideal list of compounds that were just taken off the vendor’s shelf, or are offered at unacceptable prices, makes no sense. GTM-based selection is fast, powerful in terms of coverage control.

Mean class landscapes, denoting the core as the “blue” class 2 and the remainder of the parent library as “red” class 1—mandatorily using Bayesian weighting, as class 2 is by definition a minority—are perfect indicators of the representativity of the core. At perfect *pro rata* sampling, and after compensation of subset sample sizes, a homogeneously yellow landscape, corresponding to $\bar{P}_k = 1.5 \forall k$, should be obtained. That signals the fact that, at any point of chemical space, the core subset molecules reflect the original compound density of the parent library, being neither oversampled (blue spots), nor undersampled (red spots). Figure 12 represents such mean class landscapes, obtained by (above) random drawing and (below) *pro rata* sampling of RPs of cores representing—from left to right—50, 10, 1, 0.1 and 0.01% of ChEMBL.

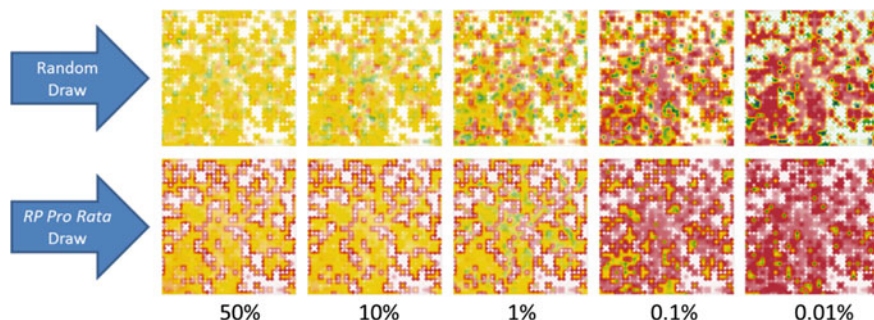


Fig. 12 Mean class landscapes, with Bayesian weighting, denoting the core as the “blue” class 2 and the remainder of the parent library as “red” class 1, at decreasing core size $c\%$ (numbers below), generated either by random draw of ChEMBL compounds (*top row*), or by *pro rata* draw of compounds from every detected responsibility pattern

Clearly, one half of the 1.3M ChEMBL compounds does indeed strongly resemble the other, and even 10% of ChEMBL is still seen to represent well the remaining 90%—even without recurring to no more sophisticated subsetting than the plain random draw. With cores of 1% or less, it becomes increasingly difficult to include representatives of every chemical space zone—hence, the clear “red shift” in the upper series of landscapes. Often, randomly picked compounds may stem from a relatively thinly populated chemical space zone—within the much smaller core, their relative importance implicitly becomes very high, and they are perceived as “oversampling” their respective chemical space zones. Therefore, the above-mentioned “red shift” is accompanied by a polarization of the landscape—emergence of a few oversampled blue “islands” in the “sea” of undersampled space.

By contrast, cores produced by *pro rata* draw from every RP show the characteristic “red border” effect in the lower series of landscapes. This is an implicit consequence of the existence of many sparsely represented RPs, with less than $1/c\%$ members, which will therefore contribute *none* of their members to the selection. Even at 50%, “singleton” RPs, each associated to exactly one molecule (there are roughly 15K such patterns, out of a total of 23K distinct RPs observed for ChEMBL compounds on the given map), cannot contribute to the selection. They provide the population of the low-density “border” regions, which will not make it into the core selection—hence the observed “red border” effect. By contrast, it can be seen that selection within the zones that can be sampled at given core size is much more homogeneous—there is clearly less polarization in the series associated to *pro rata* draws.

Alternatively, one may proceed to a “flat” draw of an equal number of representatives from each of the RPs exceeding a certain population level. The left-most density landscape (a) in Fig. 13 features a ChEMBL core of 23K compounds—one representative for each of its 23K distinct responsibility patterns. This is compared to a core of similar size, obtained by random drawing—its density trace (b) can be seen to be relatively less homogeneous, and presenting some clearly highlighted diversity holes, covered by the “flat” core.

Which of *pro rata* and flat diversity selection strategies are best-suited is a context-dependent problem. The key message here is that GTMs, exploiting the RP-based default “clustering” of molecules, is perfectly operational in diversity selection, irrespective of the used approach. One may, for example, perform a flat selection but only based on RPs with a minimum level of occurrence—which can be seen as a hybrid *pro rata*/flat approach. Such could be very useful if one wishes to maximize coverage all while ensuring that selected compounds are no singletons—i.e., close analogues thereof are available, in order to support a quick harvesting of structure-activity data after primary hit confirmation. Furthermore, diversity sampling may well be associated to already known structure-activity data or any other filters for “interesting” chemical space zones. As shown in the previous chapter, library comparison can be biased towards specific chemical space

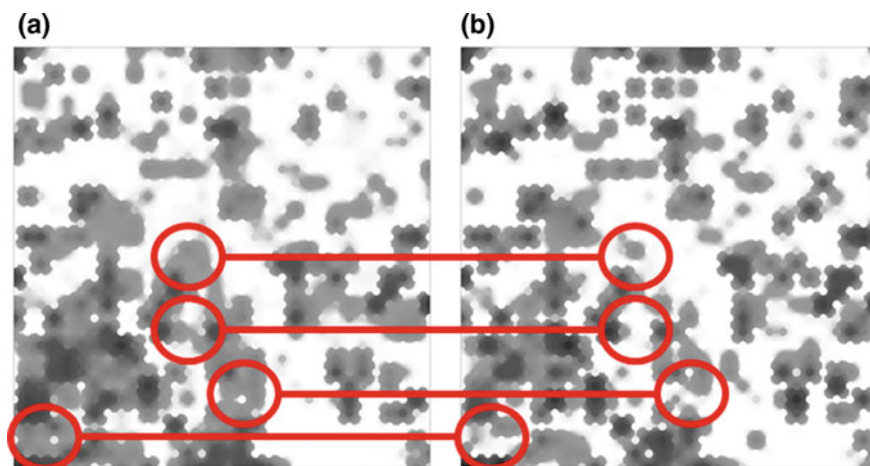


Fig. 13 **a** Density landscape of the “flat” ChEMBL core, featuring one randomly picked representative for each of the detected 23K distinct responsibility patterns on this map (same as described in Fig. 3). **b** Random drawn core of equivalent size ($\sim 1.8\%$ of ChEMBL). Connectors highlight diversity holes of the latter, covered by the “flat” selection

zones—or, diversity selection is just an application of library comparison. Or, a key advantage of a GTM is the ability to validate the proposed map, in terms of its propensity to discriminate actives from inactives, and to quantitatively predict molecular properties. A map shown to be a competent support for classification and regression models is therefore compliant with the molecular similarity principle and proposes a chemically meaningful “image” of chemical space. As such, diversity selections based on this map are also likely to fulfill the expectation of picking all the “iconic” distinct chemotypes or pharmacophores. By default, diversity selection is tributary to the initial choice of molecular descriptors, dissimilarity metric, etc. Whatever those choices, a diversity selection will emerge—and it will heavily depend on those choices. Or, as already discussed, it is very difficult to establish any “objective” quality criteria for a diversity selection aimed at designing a general-purpose screening library. Thus, the final “verdict” about the pertinence of the diversity selection can only be given a posteriori, after experimentally screening the selected library core. Instead, if one relies on a map built and shown to be similarity-principle compliant with respect to various different biological activities, the descriptor choice and the dimensionality reduction parameters (defining the manifold) has already been done and validated on the basis of quantitative statistical criteria of predictive models. If the library to be sampled is seen to fall within the Applicability Domain of such a map, the “competence” of the map in previously tackled predictive problems may be accepted as a caution for a meaningful diversity subsetting.

3.3 *Privileged Responsibility Patterns*

Consider a specific subset l of a larger compound library L , consisting of all molecules of L that have a given property—for example, all the compounds that are associated to a biological target, or, alternatively, all the compounds found active against a given target. Suppose that, out of these “specific” molecules from l , there is a fraction $f_i(RP)$ of compounds representing a given Responsibility Pattern RP —according to a given GTM model. Let, by default, the baseline occurrence rate of this RP , represent $f_L(RP)$, the overall fraction of the RP -matching molecules over the parent library L . If L is a large compound collection, representing a significant sample of the so-far synthesized and tested organic compounds, then any RP found to occur much more often in l , i.e., $f_i(RP) \gg f_L(RP)$ can be considered as *privileged* within l . A privilege score

$$\pi = f_i(RP) / f_L(RP)$$

may thus be defined. Since l is defined in terms of a specific property shared by its members, it is straightforward to link this privileged status to the property. Of course, correlation never implies causality (Horvath 2010), but it is tempting for medicinal chemists to “relate” a given pattern to a given activity. If, for example, every second active is seen to match that pattern ($f_i = 0.5$), whereas the same pattern is being encountered in only one commercial compound out of 100 ($f_L = 0.01$), this provides a rationale to specifically design and synthesize more molecules containing the pattern. The patterns which medicinal chemists love to monitor are scaffolds—hence, the “privileged scaffold” (Evans et al. 1988; Kubinyi 2006) paradigm, a very popular pedagogical method aimed at systematizing the relationships between scaffolds and therapeutic classes. Yet, it cannot be taken as granted that the best structural motif to analyze is, indeed, a single scaffold—specific, non-cyclic fragments, or scaffold families, or pharmacophores may also have a “privileged” status. The advantage of exploiting RPs in the quest for privileged patterns is that mapping of a compound on a GTM automatically defines its RP , which can be a posteriori related to the underlying structural motif (as already discussed; see the chapter introducing the RP concept).

In a previous publication (Klimenko et al. 2016), we exemplified the detection of RPs preferentially appearing within compound sets of confirmed antiviral properties and traced these RPs back to the underlying specific structural motifs. In some cases, the underlying structural motif shared by all compounds of a given privileged RP happened to be indeed a “privileged scaffold”. More often, this was not the case— RP members could alternatively share much fuzzier common structural traits (many ATP mimics featuring an anion-linker chain-heterocycle “pharmacophore-like” pattern were, for example, regrouped under a common RP). The opposite was also observed: RPs based on a well-defined scaffold with specific substitution patterns at specific points.

In the following example of privileged RP analysis, the ChEMBL database will be used as the “baseline” library L with respect to which default occurrences $f_L(RP)$, of the RPs from the previously introduced “Universal” GTM (Fig. 3), will be defined. Subsets of ChEMBL compounds associated to—being tested on—a given human biological target T from ChEMBL were used as “property-specific” subsets l , for each target with more than 500 associated compounds. Thus, the “property” that all members of a subset l have in common is *not* their strong affinity for a given target T , but the fact that they were tested on target T , irrespective of the result. This may seem odd, but the shared feature providing a common identity to all members of a subset l is the fact that they were all considered—rightly or wrongly—being *worth* testing on target T according to experts in the field. Therefore, the privileged RPs highlighted here are not the RPs privileged by the target—that is, the RPs seen to significantly enhance the change to obtain an active on that target—but rather the RPs privileged by the know-how of medicinal chemists, *believed* by medicinal chemists to relate to a given target. This analysis is therefore no rigorous structure-activity relationship, but rather a trend analysis of the human factor in drug design. An alternative analysis—in terms of rigorous measured activities—could be performed as well and, if the medicinal chemists’ flair was correct, it should conclude that patterns privileged by the target are the same as the ones privileged by chemists. On the contrary, if a target has been subjected to “carpet bombing” by High-Throughput Screening of random libraries, no privileged RPs should emerge at all, since there was little or no know-how used to associate those randomly picked screening candidates to a target.

The privilege score π has been calculated for each of the RPs, over all considered targets. Figure 14 locates on the map the five RPs (five nodes, as it turned out that all concerned RPs were single-node) with the absolute largest π scores, all targets confounded. Each of these RP is matched by compound sets of rather modest size (between 131 and 735 compounds) and “snapshots” of representative compounds are shown.

In red, the node reaching the absolute highest privilege factor corresponds to a structurally homogeneous series. This series is, strictly speaking, not based on a single “privileged” scaffold defined as a single cyclic moiety, but on an expanded aryl-oxadiazole-cyclohexyl core, with heteroatoms allowed in different positions of the aryl and cyclohexyl moieties. Such compounds are encountered within the set of compounds associated to SMO, the “smoothened” frizzled GPCR with a frequency 1757-fold higher than the default one in the entire ChEMBL database. Out of the 517 molecules associated to SMO in ChEMBL, 92 are representatives of the “red” RP, which gathers 131 compounds in its associated cluster. The other target having a still significant 12-fold enrichment of compounds from this class within the set of associated molecules is the ion channel HERG. Note that GPCRs and ionic channels are expected to privilege the same structural patterns, as many ligands binding to macromolecules of both classes are known.

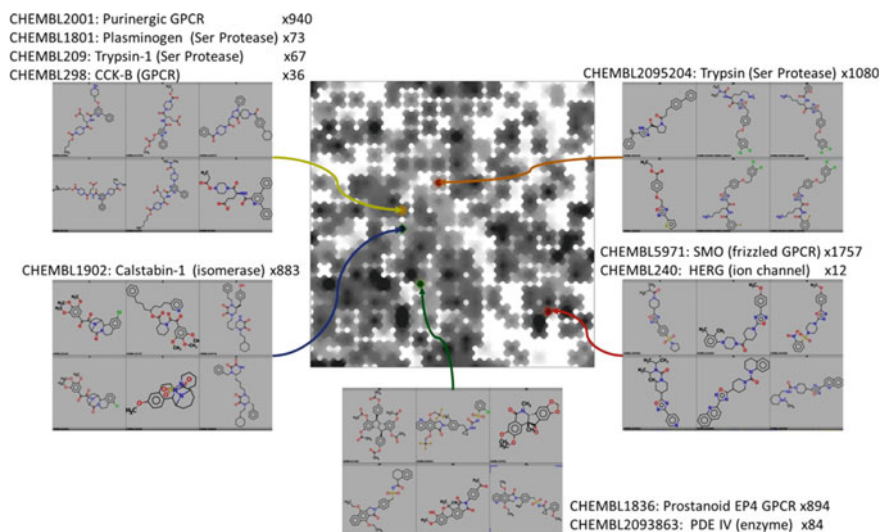


Fig. 14 Location, on the ChEMBL map (see Fig. 3 legend) of five RPs (all single-node) with top privilege scores, as colored nodes against the *grey* density plot of the entire ChEMBL. Representative samples of compounds matching each RP are shown. Next to the associated structure tables, the target or targets that are privileging each RP are denoted, next to the actual privilege score π of the RP (listed as “ $\times \pi$ ”) with respect to the target

In orange, the RP privileged by Trypsin, with a factor of 1080, matches a series of artificial peptidomimetics: rather linear compounds, with at least three aromatics connected by flexible linkers (it is worth noting the high occurrence of oxadiazole rings, though not in the same context as in SMO binders above), and often seen to embed actual amino-acids (proline, lysine) next to the non-peptidic moieties. In view of the fact that the chosen target is a protease, this makes perfect sense.

In yellow, compounds matching the third RP are even more strikingly peptide-like, consisting of several small (artificial, or amino-acid) building blocks interconnected by amide bonds. The often recurring amino-acid is glutamate, bringing a net negative charge to the species. This RP is privileged by nucleoside and peptide-binding GPCRs, and—again—proteases. Thus, the featured GTM possesses at least two—not very remote—zones dedicated to “peptide-like” molecules, and unsurprisingly associated to proteases.

Further privileged RPs cover complex patterns evoking natural product derivatives, and the analysis could be pursued for each of the significantly populated RPs (there are 504 represented by more than 100 compounds each, in the ChEMBL projection on the current map). A GTM may be manually annotated with respect to targets privileging each RP—and, as a direct consequence, compounds matching a privileged RP but not yet tested on a target are candidates of choice for further testing.

4 Conclusions

After briefly revisiting the principles of Generative Topographic Mapping as a dimensionality reduction tool in chemoinformatics, this chapter specifically focuses on the applications of GTMs for the analysis of chemical space. The key feature that dramatically enhances the analysis of chemical space through the prism of a GTM is rendering of compounds by their responsibility vectors, representing fuzzy, real-value probabilities of residence of a compound on every GTM node. Whereas on a Kohonen map, the statement “compound n resides in node k ” is either correct or false, following binary logics, on a GTM the fuzzy truth value of the above statement is nothing but the responsibility value R_{kn} . Therefore, at equal number of nodes, a GTM is much more information-rich than a Kohonen map. Albeit the latter appears to be better suited as a compound clustering tool—all residents of a node belong to the same cluster—it was shown that “binning” responsibility values can be straightforwardly used to convert this real-value vector to a short Responsibility Pattern (RP). A RP represents the non-zero responsibility values after binning, in conjunction to the node numbers to which they pertain, under the compact form of a string, or label, and may as readily serve as a clustering criterion as the Kohonen number: all compounds matching a common RP label will be regarded as members of a same cluster. In the—rather often occurring—situations of a responsibility vector dominated by a single node, the associated “single-node” responsibility pattern is formally identical to the node number identifier in the Kohonen scenario.

The fuzzy nature of GTMs versus the binary nature of Kohonen maps, and the therefrom emerging ability of the former to accommodate a much larger number of RPs at given number of nodes, will have a direct impact on the quality (structural coherence) of the clusters defined by RPs. It is well known that some of the Kohonen “garbage” nodes will “specialize” in accommodating items which do not fit into any other nodes—but need to be mapped somewhere, nevertheless. By contrast, in GTMs, such “exotic” compounds tending to be far away from the manifold in the initial descriptor space will typically be assigned, fuzzily, to many different nodes, so that single-node RPs will, in general, tend to regroup items which actually show some significant, common, structural pattern. The more populous an RP, the more difficult it is statistically to ensure that the entire set of acquired compounds is structurally homogeneous. It was found that, out of the three most densely populated single-node RPs in ChEMBL (all three being “borderline” nodes at the map edge) only one could be tentatively labeled as “garbage” node—the others are preferentially populated by Michael acceptors and oxyanionic compounds, respectively. This issue also illustrates that the nature of the “significant, common, structural pattern” assembling the compounds under a same RP label is open-ended and self-adaptive: it may be a substructure (but not necessarily a ring scaffold, as put forward by medicinal chemists), a set of related substructures, a common pharmacophore and, perhaps, even less precisely defined, a size constraint. Actually, the members of the node tentatively discarded as “garbage” do have

something in common: their size, closer to the one of typical fragments (in Fragment-Based Drug Design) than to actual drug molecules.

Albeit property prediction with GTMs is not the main topic of this contribution, this very important issue has nevertheless been discussed. First, the fact that a map can be shown to support quantitative of class-based predictive modeling provides a rigorous quality assessment of the map, something which is not provided by its other applications, such as visualization and library comparison. Second, library comparison and diversity selection—or any other form of chemical space analysis—will benefit from the knowledge contained in “property landscapes” obtained by coloring the map with diverse structure-property data. Property prediction with GTM also provided the occasion to discuss the matters of compound density, coherence of mapped properties and, in general, Applicability Domain-related issues with GTMs.

Next, the problematics of library comparison with GTMs has been revisited, on the basis of class landscapes comparing the ChEMBL collection to a roughly equally-sized set of commercial compounds. It was shown how class landscapes can be used to rapidly identify “unbalanced” zones, dominated by either of the compound collections. Reading such landscapes in parallel to property landscapes allows an immediate estimate of the properties of molecules in the unbalanced chemical space zones. Eventually, any third-party compound set—here, ChEMBL subsets from the GPCR and Kinase SARfari projects, respectively—can be used as a filter, specifically focusing the comparison of the two libraries onto chemical space zones deemed “of interest” for the ongoing research project.

Further on, the usage of GTMs as both a driver and a post hoc analyzer of diverse subset selection applications is explored. It is shown why relying on RPs to conduct diverse subset sampling is very much faster than classical methods requiring the estimation of the full dissimilarity matrix between all compounds.

Eventually, one simple but effective way to link chemical space to biological activities is discussed: Privileged Responsibility Patterns. Following the now classical “privileged scaffold” concept in medicinal chemistry, this approach has the merit of straightforward generation of RPs by mapping a library onto a “meaningful” GTM (as suggested by previous predictive challenges). It is straightforward to check whether a RP is “privileged” with respect to a given property—in the sense that its occurrence rate within compounds having that property is much larger than its occurrence rate throughout the parent library. If so, visual inspection of compounds matching the RP often suffices to find the underlying structural motif behind that RP. Therefore, since the RP is “privileged”, the underlying structural motif automatically inherits the “privileged” status and, as already highlighted, this motif does not *have* to be a privileged scaffold. The examples of the top most privileged RPs lead to the discovery of various privileged structural motifs, some being rather well-defined structural constraints (the aryl-oxadiazol-cyclohexyl moiety), while others are fuzzy, yet chemically meaningful motifs, such as “peptidomimetics”. It would have been impossible to a priori guess the peculiar motifs that should be tested for privileged status. With a chemically meaningful GTM, such guesswork is not necessary: RPs are naturally emerging hypotheses to regroup compounds

together, and the key structural motifs behind each such cluster can very often be found.

We hope this brief overview has convinced the reader of the significant strengths of GTMs in chemical space navigation and analysis.

References

- Agrafiotis, D. K. (1997). Stochastic algorithms for maximizing molecular diversity. *Journal of Chemical Information and Computer Sciences*, *37*, 841–851.
- Agrafiotis, D. K. (2003). Stochastic proximity embedding. *Journal of Computational Chemistry*, *24*, 1215–1221.
- Agrafiotis, D. K., Rassokhin, D. N., & Lobanov, V. S. (2001). Multidimensional scaling and visualization of large molecular similarity tables. *Journal of Computational Chemistry*, *22*, 488–500.
- Baell, J., & Walters, M. A. (2014). Chemical con artists foil drug discovery. *Nature*, *513*, 481–483.
- Bishop, C. M., Svensén, M., & Williams, C. K. (1998a). GTM: The generative topographic mapping. *Neural Computation*, *10*, 215–234.
- Bishop, C. M., Svensén, M., & Williams, C. K. I. (1998b). Developments of the generative topographic mapping. *Neurocomputing*, *21*, 203–224.
- ChemAxon. (2007). *Fingerprint and descriptor generation—GenerateMD*. Budapest. Retrieved September, 2016, from <https://docs.chemaxon.com/display/docs/163210/Fingerprint+and+descriptor+generation+-+GenerateMD>.
- Dahlin, J. L., Nissink, J. W. M., Strasser, J. M., Francis, S., Higgins, L., Zhou, H., et al. (2015). PAINS in the assay: Chemical mechanisms of assay interference and promiscuous enzymatic inhibition observed during a sulfhydryl-scavenging HTS. *Journal of Medicinal Chemistry*, *58*, 2091–2113.
- Dunteman, G. H. (1989). *Principal components analysis*. : Sage Publications.
- Evans, B. E., Rittle, K. E., Bock, M. G., Dipardo, R. M., Freidinger, R. M., Whitter, W. L., et al. (1988). Methods for drug discovery: Development of potent, selective, orally effective cholecystokinin antagonists. *Journal of Medicinal Chemistry*, *31*, 2235–2246.
- Gaspar, H. A., Baskin, I. I., Marcou, G., Horvath, D., & Varnek, A. (2014). Chemical data visualization and analysis with incremental generative topographic mapping: big data challenge. *Journal of Chemical Information and Modeling*, *55*, 84–94.
- Gaspar, H. A., Baskin, I. I., Marcou, G., Horvath, D., & Varnek, A. (2015). GTM-based QSAR models and their applicability domains. *Molecular Informatics*, *34*, 348–356.
- Gaspar, H., Marcou, G., Horvath, D., Arault, A., Lozano, S., Vayer, P., et al. (2013). Generative topographic mapping-based classification models and their applicability domain: Application to the biopharmaceutics drug disposition classification system (BDDCS). *Journal of Chemical Information and Modeling*, *53*, 3318–3325.
- Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., et al. (2011). ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, *40*, D1100–D1107.
- Horvath, D. (2010). Quantitative structure-activity relationships: In silico chemistry or high tech alchemy? *Revue Roumaine de Chimie*, *55*, 783–801.
- Horvath, D., & Barbosa, F. (2004). Neighborhood behavior—The relation between chemical similarity and property similarity. *Current Trends in Medicinal Chemistry*, *4*, 589–600.
- Horvath, D., Lisurek, M., Rupp, B., Kühne, R., Specker, E., Von kries, J., et al. (2014). Design of a general-purpose European compound screening library for EU-OPENSREEN. *ChemMed-Chem*, *9*, 2309–2326.

- Horvath, D., Marcou, G., & Varnek, A. (2009). Predicting the predictability: A unified approach to the applicability domain problem of QSAR models. *Journal of Chemical Information and Modeling*, *49*, 1762–1776.
- Horvath, D., Marcou, G., & Varnek, A. (2013). Do not hesitate to use Tversky-and other hints for successful active analogue searches with feature count descriptors. *Journal of Chemical Information and Modeling*, *53*, 1543–1562.
- Kireeva, N., Baskin, I., Gaspar, H. A., Horvath, D., Marcou, G., & Varnek, A. (2012). Generative topographic mapping (GTM): universal tool for data visualization, structure-activity modeling and dataset comparison. *Molecular Informatics*, *31*, 301–312.
- Klimenko, K., Marcou, G., Horvath, D., & Varnek, A. (2016). Chemical space mapping and structure-activity analysis of the ChEMBL antiviral compound set. *Journal of Chemical Information and Modeling*, *56*, 1438–1454.
- Kohonen, T. (1984). *Self-organization and associative memory*. Heidelberg: Springer.
- Kohonen, T. (2001). *Self-organizing maps*. Heidelberg, Berlin, Germany: Springer.
- Kubinyi, H. (2006). Privileged structures and analogue-based drug discovery. In J. G. R. Fischer (Ed.), *Analogue-based drug discovery*.
- Johnson, M., Basak, S., & Maggiora, G. (1988). A characterization of molecular similarity methods for property prediction. *Mathematical and Computer Modelling*, *11*, 630–634.
- Johnson, M., & Maggiora, G. M. (1990). *Concepts and applications of molecular similarity*. New York: Wiley.
- Maldonado, A. G., Doucet, J. P., Petitjean, M., Fan, B. T. (2006). Molecular similarity and diversity in chemoinformatics: From theory to applications. *Molecular Diversity*, *10*, 39–79.
- Papadatos, G., Cooper, A. W. J., Kadiramanathan, V., Macdonald, S. J. F., McLay, I. M., Pickett, S. D., et al. (2009). Analysis of neighborhood behavior in lead optimization and array design. *Journal of Chemical Information and Modeling*, *49*, 195–208.
- Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D., & Weinberger, L. E. (1996). Neighborhood behavior: A useful concept for validation of “molecular diversity” descriptors. *Journal of Medicinal Chemistry*, *39*, 3049–3059.
- Ruggiu, F., Marcou, G., Varnek, A., & Horvath, D. (2010). Isida property-labelled fragment descriptors. *Molecular Informatics*, *29*, 855–868.
- Sidorov, P., Gaspar, H., Marcou, G., Varnek, A., & Horvath, D. (2015). Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds. *Journal of Computer-Aided Molecular Design*, *29*, 1087–1108.
- Sushko, I., Novotarskyi, S., Korner, R., Pandey, A. K., Cherkasov, A., Lo, J. Z., et al. (2010). Applicability domains for classification problems: Benchmarking of distance to models for Ames mutagenicity set. *Journal of Chemical Information and Modeling*, *50*, 2094–2111.
- Tetko, I. V., Sushko, I., Pandey, A. K., Zhu, H., Tropsha, A., Papa, E., et al. (2008). Critical assessment of QSAR models of environmental toxicity against tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection. *Journal of Chemical Information and Modeling*, *48*, 1733–1746.
- Turner, D. B., Tyrrell, S. M., & Willett, P. (1997). Rapid quantification of molecular diversity for selective database acquisition. *Journal of Chemical Information and Modeling*, *37*, 18–22.

Part III

Applications

On Applications of QSARs in Food and Agricultural Sciences: History and Critical Review of Recent Developments

Supratik Kar, Kunal Roy and Jerzy Leszczynski

Abstract During the past decade, a large number of reports described the roles of *in silico* approaches in the development of new molecules in the field of pharmaceuticals, agrochemicals, food science, materials science, environmental science, etc. *In silico* techniques like quantitative structure-activity relationships (QSAR), pharmacophore, docking and virtual screenings are playing crucial roles for the design of “better” molecules that may later be synthesized and assayed. This chapter presents the currently available information on diverse groups of molecules with applications in agriculture and food science that have been subjected to *in silico* studies. A hefty numbers of successful applications of QSARs in the development of agrochemicals, food products and food supplements are thoroughly discussed. The QSAR studies summarized here would help readers to understand the proper mechanism for the activity of miscellaneous agrochemicals and food products as well as the interaction between the free radicals and antioxidant molecules. This chapter justifies the need to develop additional QSAR models in combination with other *in silico* approaches for the design of *better* agrochemicals, food and food supplements, especially antioxidants and flavoring agents, in order to explore the largely unexplored field of plant sources in addition to synthetic molecules as well as to reduce time and cost involvement in such exercises. Further, we have enlisted most of the available agrochemical, food and flavor databases for convenience of researchers working in the area along with an extensive list of software tools.

Keywords Agriculture • Agrochemicals • Antioxidant • Food • Food supplements • Growth regulators • Phytochemicals • QSAR

S. Kar • J. Leszczynski (✉)

Department of Chemistry and Biochemistry, Interdisciplinary Nanotoxicity Center,
Jackson State University, Jackson, MS, USA

e-mail: jerzy@icnanotox.org

K. Roy

Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical
Technology, Jadavpur University, Kolkata 700032, India

© Springer International Publishing AG 2017

K. Roy (ed.), *Advances in QSAR Modeling*, Challenges and Advances

in Computational Chemistry and Physics 24, DOI 10.1007/978-3-319-56850-8_7

1 Introduction

Agriculture has played a crucial role in the development of human civilization. It is widely believed that domestication of plants (and animals) allowed humans to settle in a place and give up their previous migrant life style. Until the industrial revolution, the majority of the human population labored in agriculture. Development of agricultural techniques and constituents of fertilizers has steadily increased agricultural productivity. However, the agricultural productivity may be greatly diminished by certain threats like weeds, fungus, pests, insects etc. Additionally, weakened soil fertility and insufficient plant growth regulators affect the productivity many folds with the flow of time. To overcome the devastating effects of those threats, the use of agrochemicals has been increasing in the cultivation fields (Lamberth et al. 2013) Therefore, there is a consistent requirement to search for efficient agrochemicals with enhanced productivity and lower health hazards.

In the food science, development of better food, food supplements and nutraceuticals are always a challenge. Employing phytochemicals and their structures, a large number of modifications can be performed using *in silico* approaches. For a long time, a number of reports have suggested the use of computer models for the development of new food materials. Again, food proteins contain peptide sequences that positively affect specific health markers as shown by different *in vitro* bioassays. Such peptides, which are called bioactive peptides (BAPs), may be released from a wide range of dietary proteins. The efficacy and activity of BAPs have been studied *in vivo*, suggesting in certain instances that food protein-derived BAPs may be relevant to human health (Le Maux et al. 2015). An increasing number of peptide sequences identified in various food protein hydrolysates have been reported in the literature over the past few years. The technology used to identify these sequences within food protein hydrolysates or their fractions has greatly evolved. Nevertheless, challenges such as the reliable identification of short peptide sequences are still an issue for a more comprehensive understanding of dietary BAP structures where molecular modelling may be effective (Le Maux et al. 2015).

Fragrance and flavor substances are strong-smelling organic compounds. Their major common characteristic is a pleasant odor and/or a pleasant taste. A fragrance substance is used as a component in a perfume or a perfumed product, while a flavor substance is used to enhance the flavor of beverages and food products. The development and search for new flavoring agent is complicated. There are multiple factors that affect flavoring efficacy; e.g., solubility, stability at wide pH and temperature ranges, clean specific taste without post-flavor effects, and finally, the most important factor is the safety of human health. Because of all these factors, there is a clear advantage to develop *in silico* models in order to understand the mechanism and to use these models to develop and synthesize new potent flavoring agents (Zhong et al. 2013).

Antioxidants have occupied a larger area of study in the food science due to availability of wide range of chemical classes of antioxidants. The antioxidants have enormous biological significance against an array of deadly diseases caused

due to systemic free radical overload. Smoking, pollution, consumption of junk food etc. result in continuous production and accumulation of the deadly free radicals within the human system. The free radicals are constitutively produced within the human system during various physiological processes by enzymatic and non enzymatic reactions. Within the human system, the free radicals are efficiently neutralized by a series of antioxidant enzymes, maintaining a balance between free radical production and destruction. However, during free radical overload, these systemic antioxidants fail to defend the system, leading to oxidative stress. When there is a large increase in amount of oxidants, damage to normal cells and tissues occur with DNA being the prime target. Thus, attempts have been made to restore normal oxidative balance either by the therapeutic application of the endogenous antioxidants or by using antioxidant rich foods/drugs. The huge importance of antioxidants has lead researchers to search for synthetic antioxidants with improved activity and reduced toxicity. Thus, a great deal of recent research has been concentrated on the design and synthesis of antioxidants (Roy and Mitra 2009).

Computer-aided molecular design has been generally accepted and extensively applied in the area of modern drug discovery, ecotoxicological modeling and design of agrochemicals for its high efficiency in the design of new compounds and optimization of lead compounds, saving both time and economic costs in the large-scale experimental synthesis and biological tests. Quantitative structure-activity relationship (QSAR) helps us to understand structure-activity relationship (SAR) in a quantitative manner. It is one of the most important applications of chemometrics, giving information useful for the design of new compounds acting on a specific target. The QSAR attempts to find a consistent relationship between biological activity or toxicity and molecular properties. Thus, QSAR models can be used to predict the activity of more advanced active agrochemicals. Not only in the field of agrochemicals, QSAR had also an enormous role in the food industry. The QSAR enables identification of the response pharmacophore as well as the essential molecular fragments imparting antioxidant propensity to various classes of chemicals, and serves as a reliable tool for searching efficient antioxidant molecules with improved activity. Furthermore, several examples of QSAR are found for the study of food protein-derived BAPs. Although a hefty number of QSAR models have been reported for the antioxidant activity in the last decade, still a lot of new QSAR studies are required in the area of food flavor, taste and food supplements for advancement of food industries to the next advanced level (Roy et al. 2015a, b).

2 Agriculture

With the modern civilization, new agricultural techniques and effective fertilizers have steadily increased agricultural productivity. Whatever man has reaped must be well protected. At the same time, future yields must be improved. Unfortunately this is not so, because between the times a crop is cultivated and consumed by man, considerable quantity of agricultural product is wasted or destroyed by certain

threats like weeds, fungus, insects etc. However, the agricultural productivity may be greatly diminished by those threats. Therefore, before creating the solutions for the mentioned threats, one has to know the threat types and their diversity along with the extent of effects. The harmful effects of the mentioned threats have been discussed here.

2.1 Weeds

In 1967, the Weed Science Society of America defined a weed as “*a plant growing where it is not desired*” (Buchholtz 1967). In 1989 the Society’s definition was changed to define a weed as “*any plant that is objectionable or interferes with the activities or welfare of man*” (Vencill 2002). Weeds are generally considered as unwanted plants in human made settings like agricultural areas, gardens etc. because (1) they might restrict light to the desirable plants, (2) they can take the nutrients from soil leaving the desired plant unfed and making them less productive, (3) they can spread plant pathogens that infect and diminish the quality of crop.

2.2 Fungus

A fungus is a eukaryotic organism that is a member of the kingdom “Fungi”. Preharvest losses due to fungal diseases in world crop production may amount to 12% or even more in developing countries (Hartman et al. 2004). Phytopathogenic fungi and their harmful effects are discussed in Table 1.

2.3 Insects

Insects are the biggest class of arthropods. They are the most diverse group of animals on the planet. They are most diverse at the equator and their diversity declines toward the poles. There has been an unceasing struggle between man and his insect enemies to protect the agricultural outcomes. Numerous advances have been made by man in evolving newer and deadlier weapons to fight the war against insects. We have cited a few instances of harmful effects of insects in agricultural field in Table 1.

2.4 Viruses

Plant viruses have limited ability to enter intact host cells and they mainly depend on insect, mite, nematode, or fungus vectors to gain entry. Viruses usually invade

Table 1 Occurrence and disease caused to plants by different agriculture threats

	Disease type/occurrence
Fungi	
<i>Botrytis</i>	A dark to light brown rot forms in the diseased tissue
<i>Thielaviopsis basicola</i>	Cause root rot and branch dieback on a number of woody and herbaceous plants including holly, begonia, geranium, poinsettia etc.
<i>Botryosphaeria</i>	Cause a branch dieback on horse chest nut, tulip poplar, crabapple, pine, oak etc. Cankers girdle and kill twigs and branches
<i>Cylindrocladium</i>	Attacks azaleas, rhododendrons, camellias, junipers, white pine and involved in damping-off, wilt, root rot, stem canker, crown rot etc.
<i>Phytophthora</i>	Cause root rot of herbaceous and woody ornamentals including arborvitae, azalea, dogwood etc., and cause late blight in Potato
<i>Pythium</i>	Three common pythium species are <i>Pythium irregulare</i> , <i>Pythium aphanidermatum</i> and <i>Pythium ultimum</i> causes root rot disease
<i>Rhizoctonia</i>	The pathogenic fungus known to cause root rots, stem rots, damping-off, and in some cases, a blight of leaves of several plants
<i>Verticillium</i>	Cause disease called Verticillium wilt. The fungus plugs the water conducting vessels thus restricting flow to branches and leaves
<i>Oidium</i>	Cause powdery mildew in grapes. White, mealy fungal growth develops on leaves, flowers, and stems
<i>Alternaria</i>	Cause alternaria leaf spot. Small purplish spots form on leaves. Their centers become brown while the leaf yellows
<i>Fusarium</i>	Causes wilt disease followed by blockage and breakdown of xylem, symptoms appear in plants such as leaf wilting, yellowing and death
<i>Uromyces</i>	Small blisters containing rust-red spores form on leaves.
<i>Phoma</i>	Cause blight in Vinca, Beet, Sweet potato etc. Initially branches become dark followed by the entire plant blackens at the base and dies
<i>Venturia</i>	Cause blight or scab in Willow tree
<i>Magnaporthe</i>	Infect a number of important cereals including rice, wheat, rye, barley, and pearl millet causing diseases called blast disease or blight disease
<i>Phakopsora</i>	Causes soybean rust that affects soybeans and other legumes
<i>Puccinia</i>	Causal agents of severe rusts of virtually all cereal grains and grasses

(continued)

Table 1 (continued)

	Disease type/occurrence
<i>Septoria</i>	Causes leaf spot mainly in marigold. Oval to irregular gray to black spots with tiny dots peppering their surface
Insects	
<i>Aphids</i>	Cause injury to plants by sucking the sap and juices from the soft, new growth. Affect the plant to lose vigor, wilt, distort or show spots
<i>Earwigs</i>	Primarily scavengers of dead insects and rotted plant materials but may also feed on live plants. Targets are marigolds, dahlias, zinnias, roses etc.
<i>Geranium budworm</i>	Young larvae of this insect tunnel into small flower buds, while larger caterpillars eat flower petals, chewing the reproductive flower parts
<i>Spider Mites</i>	Produce lesions on the green epidermal cells and thus can significantly reduce the photosynthetic capability of plants
<i>Thrips</i>	They feed by sucking juices from the plant causing stippling, or small scars, on leaves, flowers and fruit
<i>Flea beetles</i>	Chew small pits into leaves, giving the appearance that they may have been blasted with fine shot. Common on cabbages, tomatoes, and beans
<i>Elude corn seed maggot</i>	Attacks the seeds of many warm-season vegetables (bean, corn, melon, cucumber) planted early when soil is still cool and often damp
<i>Spinach Leaf Miners</i>	Tunneling within the leaves especially of spinach, beets and chard
<i>Squash bugs</i>	Large areas of the plant become girdled and wilt
<i>Hornworms</i>	Quickly defoliate tomatoes, potatoes, eggplants, peppers and green fruit
<i>Psyllids</i>	The insect causes tomato to stop forming or ripening, and many small potato tubers to develop and sprout prematurely before harvest
<i>Rice hispa</i>	Cause scraping of the upper surface of the leaf blade leaving only the lower epidermis as white streaks parallel to the midrib. Tunneling of larvae through leaf tissue causes irregular translucent white patches
<i>Stem Borer</i>	The drying of growing part of the plant is known as 'dead heart'. Dead heart is created in early life of the plant before flowering and 'White head' occurs at flowering resulting in drying of the entire panicle
<i>Army worms</i>	Leaf feeding and cut the wheat heads from the plant stem
<i>Green bug</i>	They cause distinct damage by yellowing of leaves and the occurrence of chlorotic spots (wheat)

(continued)

Table 1 (continued)

	Disease type/occurrence
<i>Scale insects</i>	Found on stems and the undersides of leaves but can be on top of the leaves. Scales suck the juice of plants, stunting the plants growth
<i>Potato beetle</i>	Feed on foliage and may completely eliminate the crop
<i>Boll weevil</i>	Destructive cotton pest which feeds on cotton buds and flowers
<i>Corn rootworm</i>	Damage caused by larval feeding of the insect
<i>Rose chafer</i>	The adult beetle feeds on the foliage, flowers, and fruit of plants
<i>Asian longhorned beetle</i>	During the larval stage, this insect bores deep into a tree's heartwood causes tunneling damages and eventually kills the tree
<i>Citrus longhorned beetle</i>	They chew into the tree, forming a tunnel and death of the plant (citrus, pecan, apple etc.)
<i>Bark beetle</i>	Kill pine tree by boring through the bark into the phloem layer
Bacteria related to necrotic diseases	
<i>P. syringae</i> pv. <i>Tabaci</i>	Wildfire of tobacco
<i>P. syringae</i> pv. <i>Lachrymans</i>	Angular leaf spot of cucumber
<i>X. translucens</i>	Bacterial streak of barley and black chaff of wheat
<i>X. oryzae</i> pv. <i>Oryzae</i> and <i>X. oryzae</i> pv. <i>Oryzicola</i>	Bacterial leaf blight and leaf streak of rice
<i>X. vesicatoria</i> and <i>P. syringae</i> pv.	Bacterial spot and bacterial speck of tomato
<i>P. syringae</i> pv. <i>Syringae</i> , <i>X. phaseoli</i> and <i>P. syringae</i> pv. <i>Phaseolicola</i>	Bacterial brown spot, common blight and halo blight of bean
<i>Streptomyces scabies</i>	Scab of potato
Bacteria related to vascular diseases	
<i>Ralstonia solanacearum</i> race 1	Southern bacterial wilt of solanaceous plants
<i>R. solanacearum</i> race 2	Moko disease of bananas
<i>Clavibacter michiganensis</i> subsp. <i>Sepedonicus</i>	Potato ring rot
<i>C. michiganensis</i> subsp. <i>Michiganensis</i>	Bacterial canker and wilt of tomato
Bacteria related to the soft rots	
<i>Erwinia carotovora</i> subsp. <i>Atroseptica</i>	Tomato, potato
<i>Erwinia carotovora</i> subsp. <i>Betavasculorum</i>	Sugar beet, sunflower, potato
<i>E. chrysanthemi</i> pv. <i>Zea</i> .	Maize, pineapple, potato, banana
<i>E. chrysanthemi</i> pv. <i>Dianthicola</i> .	Carnation, chicory, artichoke, dahlia, tomato, potato

(continued)

Table 1 (continued)

	Disease type/occurrence
Bacteria related to tumor diseases	
<i>Agrobacterium vitis</i>	Crown gall, Grape
<i>Agrobacterium rubi</i>	Cane gall, Raspberry and blackberry
<i>Agrobacterium rhizogenes</i>	Hairy root disease, Apple
<i>Pseudomonas savastanoi</i>	Galls, Olive
<i>Pantoea agglomerans</i> pv. <i>gypsophila</i>	Crown and root gall, Gypsophila
<i>Rhizobacter dauci</i>	Carrot gall, Carrot
<i>Rhodococcus fascians</i>	Leaf gall, Nicotiana

their host plants systemically from the initial entry sites. Viral infections are generally localized in the roots, stems or leaves of the infected plant. They are mosaic causing necrosis of leaves, petioles and stem on different solanaceous plants. Tobacco mosaic virus, Cucumber mosaic virus and tobacco spot virus are observed on digitalis (the infection causes characteristic patterns on the leaves, i.e., mottling and discoloration) and a strain of cucumber mosaic virus is detected on hyoscyamus. The viruses show disease symptoms on rauwolfia, tobacco, datura, vinca and eucalyptus. Other viruses reported on different medicinal plants are yellow vein mosaic virus, graft transmissible virus, distortion mosaic virus, rugose leaf curl etc.

2.5 Bacteria

The major groups of bacteria that adversely affect plants are the bacterial plant pathogens. Although the changes in plant physiology induced by these organisms are generally considered to be detrimental to plant health, some have been exploited as favorable from a horticultural perspective. Symptoms induced by phyto pathogenic bacteria range from local areas of cell death such as leaf spots, cankers and scabby lesions, to wilts, yellowing, tissue liquefaction, and tumor formation. Plant pathogenic bacteria have been broadly classified into four classes discussed in Table 1 (Beattie 2006).

2.6 Other Pests

Rodents cause severe damage to the agricultural products. Rodents have sharp incisors that they use to gnaw wood and cause considerable spoilage to stored crop.

3 Agrochemicals

“Agrochemical”, a contraction of agricultural chemical, is a generic term for the various chemical products used in agriculture. In most cases, agrochemicals refer to the broad range of pesticides, including herbicides, fungicides and insecticides. They may also include hormones and other chemical growth agents of plant (Waxman 1998).

3.1 *Herbicides*

Herbicide comes from the Latin herba, meaning “plant” and caedere, meaning “to kill”. Therefore, herbicides are chemicals that kill plants. The definition accepted by the Weed Science Society of America (Vencill 2002) is that a herbicide is “a chemical substance or cultured organism used to kill or suppress the growth of plants.” In effect, a herbicide disrupts the physiology of a weed over a long enough period to kill it or severely limit its growth. The classification of herbicides is cited below according to their mode of actions. They are broadly classified into seven groups in Table 2 (Zimdahl 2007).

3.2 *Fungicides*

Fungicides are chemical compounds used to kill or inhibit fungi or fungal spores. In a broader sense, fungicides either are mobile or immobile. Systemic fungicides are mobile compounds that are able to penetrate the cuticle of leaves and stems and enter the plant deeper tissues, whereas immobile fungicides reside entirely at the site of application. The classification of fungicides is sited below according to their mode of actions. They are broadly classified into nine groups in Table 3 (Copping and Hewitt 1998a).

3.3 *Insecticides*

The use of insecticides is believed to be one of the major factors behind the increase in agricultural productivity in the 20th century. Nearly all insecticides have the potential to significantly alter ecosystems; many are toxic to humans; and others are concentrated in the food chain. It is necessary to balance agricultural needs with environmental and health issues while using insecticides. The classification of insecticides is sited below according to their mode of actions in Table 4 (Copping and Hewitt 1998b).

Table 2 Classification of herbicides based on the mechanism of action (MOA) with examples

Classification	Site/type of action	Example of herbicides
<i>Inhibitors of photosynthesis</i>	Photosystem II, site A	Phenyl-Carbamates, Pyridazinone, Triazines, Chlorotriazines
	Photosystem II, site A but with different binding behavior	Amides, Urea
	Photosystem II, site B	Benzothiadiazoles, Nitriles, Phenyl-pyridazines
	Photosystem I-electron diverters	Diquat, Paraquat
<i>Inhibitors of pigment production</i>	Carotenoid biosynthesis	Amitrole
	Phytoenedesaturase with blockage of carotenoid biosynthesis	Norflurazon, Fluridone
	1-deoxy-D-xyulose 5 phosphate synthatase (DOXP synthase)	Clomazone
	Protoporphyrinogen oxidase (Protox)	Diphenylethers, Acifluorfen, Phenylthalamides, Thiadiazoles, Triazinones, Pyrazolotriazinones
<i>Fatty acid biosynthesis inhibitors</i>	Acetyl-CoA carboxylase (ACCase)	Aryloxyphenoxypropionates (Clodinafop, Fluazifop, Quizalofop) and Cyclohexanediones (Clethodim, Sethoxydim, Tralkoxydim)
	Lipid synthesis, but not by ACCase inhibition	Carbamothioates (Butylate, Cycloate, Pebulate, Triallate)
	Inhibiting biosynthesis of very long chain fatty acids	Chloroacetamides (Acetochlor, Alachlor, Dimethenamid, Flufenacet)
<i>Amino acid biosynthesis inhibitors</i>	Acetolactate synthase (ALS)-acetohydroxy acid synthase (AHAS)	Sulfonylureas (Bensulfuron, Metsulfuron, Triflursulfuron), Imidazolinones, Pyrimidinylthio-benzoate, Triazolopyrimidines
	5-enolpyruvyl-shikimate-3-phosphate synthase (EPSP)	Glyphosate
	Glutamine synthatase (GS)	Glufosinate
<i>Cell growth inhibitors</i>	Microtubule assembly	Dinitroanilines (Benfenin, Ethalfuralin, Oryzalin, Trifluralin), Terephthalic acid, Dacthal
	Mitosis	Carbetamide
	Cell wall synthesis	Nitriles (Dichlobenil), Benzamides (Isoxaben)
<i>Auxin-like action-growth regulators</i>	Synthetic auxins	Phenoxy acids, Arylaliphatic or benzoic acids, Picolinic acids, Quinolinecarboxylic acid
	Indoleacetic acid (IAA) transport	Naptalam, Diflufenzopyr
<i>Inhibition of respiration</i>	By uncoupling oxidative phosphorylation	Arsenite, Phenol

Table 3 Classification of fungicides based on the MOA with examples

Classification	Site of action	Example of some fungicides
<i>Sterol biosynthesis inhibitors</i>	Inhibition of C14 α -demethylation	1,2,4-triazoles (Triadimenol, Propiconazole), Imidazoles (Imatalil, Prochloraz), Pyrimidinylcarbinols (Fenarimol, Nuarimol), Piperazines (Triforine)
	Inhibition of $\Delta^{8,7}$ Isomerase and Δ^{14} Reductase	Fenpropimorph, Fenpropidin, Dodemorph, Spiroxamine
<i>Glycerophospholipid biosynthesis inhibitors</i>	Inhibition of phosphatidylcholine synthesis	Iprobenfos, Edifenphos, Isoprothiolane
	Inhibition of phosphatidylinositol synthesis	Validamycin A
<i>Nucleic acid biosynthesis inhibitors</i>	Inhibition of DNA synthesis	Hymexazol
	Inhibition of RNA synthesis	Acylufanines (Metalaxyl, Metalaxyl M), Hydroxypyrimidines (Ethirimol, Dimethirimol Bupirimate), Phenoxyquinolines (LY214352)
<i>Tubulin biosynthesis inhibitors</i>	Inhibition of cell division	Benzimidazoles (Benomyl, Thiabendazole, Carbendazim, Fuberidazole), Diethofencurb
<i>Chitin biosynthesis inhibitors</i>	Inhibition induces the collapse of cell wall integrity, leads to swelling and bursting of hyphal tips and spore germ tubes	Polyoxins (Polyoxin B, Polyoxerim D)
<i>Melanin biosynthesis inhibitors</i>	Inhibition of development of infection hyphae and subsequent penetration of the host epidermis	Tricyclazole, Pyroquilon, Phthalide, KTU 3616
<i>Protein biosynthesis inhibitors</i>	Inhibiting the protein synthesis	Blasticidin, Kasugamycin
<i>Respiration inhibitors</i>	Inhibition of Complex II	Carboxamides (Carboxin, Fenfuram, Methfuroxam), Thifluzamide
	Inhibition of Complex III	Famoxadone, Azoxystrobin, Kresoxim-methyl
	Inhibition of oxidative phosphorylation	Nitrophenols (Dinocap, Nitrothal-isopropyl, Fluazinam), Fentins (Fentin acetate, Fentin hydroxide)
<i>Interference with cell membrane structure</i>	Destabilizing the cell membrane	Guanidines (Dodine, Guazatine)

Table 4 Classification of insecticides based on the MOA with examples

Classification	Type of action	Example of some insecticides
<i>Inhibitors acting by insect nervous system disruption</i>	Organophosphorus Insecticides	Phosphate, Phosphonate, Phosphorothionate, Phosphorothiolate, Phosphoramidate, Phosphorothionate, Parathion, Terbufos
	Carbamate Insecticides	Methornyl, Carbofuran, Carbosulfan, Ethiofencarb, Pirimicarb, Carbaryl
	Interact with neurotransmitter ligand recognition sites	Nicotine, Nereistoxin, Bensultap, Cartap, Imidacloprid, Amitraz.
	Insecticides that interfere with ion channels	Pyrethroids (Permethrin, Deltamethrin, Tefluthrin, Silafluofen), DDT
<i>Inhibition of oxidative phosphorylation</i>	Inhibit the mitochondrial electron transport chain by binding with complex I at coenzyme site Q	Fipronil, rotenone, dinitro-o-cresol, pyrimidifen, fenazaquin
	Inhibition of respiration at complex III	Naphthoquinones

Table 5 A list of bactericides with exhaustive examples

Bactericide	Examples
Copper compounds	Ammoniacal copper sulfate, Copper oxide, Copper oxyquinolate, Copper hydroxide, Copper oxychloride, (Tri)basic copper sulphate, Copper sulphate + lime, Copper oxychloride + maneb, mancozeb or chorothalonil
Antibiotics	Kasugamycin, Oxytetracyclin, Streptomycin
Disinfectants	Acetic acid 1(M), Benzalkonium chloride, Ethanol 70% or 80%, Isopropanol 70%, Propionic acid 1(M), Quaternary ammonium compounds, Calcium hypochloride, Sodium hypochloride, Chlorine dioxide, Hydrogen peroxide with peracetic acid, Ozone, Phenolic and cresolic compounds, Formaldehyde, Potassium permanganate
Other compounds	Flumequin, Fosetyl-aluminium, 7-Chloro-1-ethyl-6-fluoro-1,4-dihydro-4-exo-3-quinoline carboxylic acid, Oxolinic acid

3.4 Bactericides

Bactericides are chemicals that prevent bacterial infections. They kill bacteria on contact and must be used before the bacteria infect a plant. Most of the bacterial pathogens are systemic, but most of the bactericides are surface protectants (Table 5). Bacteria also develop resistance to antibiotics very quickly. Persistence of antibiotics in plants is very low, requiring that antibiotics be applied once every

4–5 days. This is impractical and will be uneconomical. The current bactericides cannot reach sites where bacteria overwinter, such as blighted wood, cankers, and lesions in the case of woody plants. Only partial control of bacterial diseases is achievable with the available bactericides (Janse 2005).

3.5 *Virucides*

Chemical control of viral diseases is almost not practical. However, the vectors (certain fungi, insects etc.) of viruses can be controlled to a certain extent to reduce the disease spread. Chemical control of viral diseases is difficult to achieve. Only ribavirin has been shown to reduce virus diseases. Ribavirin is effective against Potato S virus (Carlavirus) and Odontoglossumringspot virus (Tobamovirus). Another method of controlling virus diseases by chemicals is by using plant activators, which induce systemic resistance against virus infection. Preplant application of the plant activator acibenzolar-S-methyl (Actigard) effectively controls Tomato spotted wilt virus (TSWV) in tomato (Vidhyasekaran 2004).

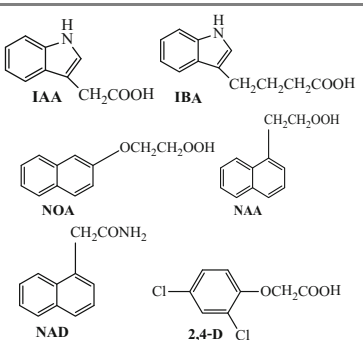
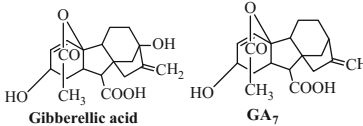
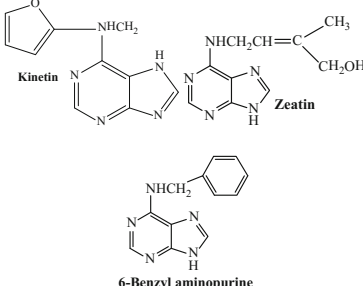
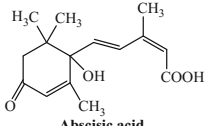
3.6 *Plant Growth Regulators*

Plant hormones (also known as phytohormones) are chemicals that regulate plant growth. Plant hormones are signal molecules produced within the plant, and occur in extremely low concentrations (Kar and Roy 2012). Hormones also determine the formation of flowers, stems, leaves, the shedding of leaves, and the development and ripening of fruit. The functions of those growth regulating agents are reported below in Table 6.

4 **Food, Food Supplements and Phytochemicals**

Food science includes the development of new food products, design of processes to produce foods, choice of packaging materials, shelf-life studies, sensory evaluation of products using panels or potential consumers, as well as microbiological and chemical testing. Food scientists may study more fundamental phenomena that are directly linked to the production of food products and their properties. Food science brings together multiple scientific disciplines such as microbiology, chemical engineering, and biochemistry to increase the production rate of food products due to the immense population growth and decrease efficiency of soil productivity (Fratamico and Bayles 2005). Food science is one of the unexplored fields in terms of employing cheminformatics for new molecule design. Although a large number of studies have been undertaken for antioxidants using in silico tools,

Table 6 A list of plant growth regulators, their chemical structure and function

Growth regulator	Chemical structure	Function
Auxins	 <p>IAA CH_2COOH IBA $\text{CH}_2\text{CH}_2\text{CH}_2\text{COOH}$</p> <p>NOA $\text{OCH}_2\text{CH}_2\text{OOH}$ NAA $\text{CH}_2\text{CH}_2\text{OOH}$</p> <p>NAD CH_2CONH_2 2,4-D Cl OCH_2COOH Cl</p>	They are involved in different growth processes in plants like internodal elongation, leaf growth, initiation of vascular tissues, cambium activity, fruit setting in absence of pollination, fruit growth, apical dominance, inhibition of root growth, inhibition of lateral buds
Gibberellins	 <p>Gibberellic acid GA_7</p>	They are involved in vegetative and fruit growth, breaking dormancy (seed germination), flower initiation and induction of parthenocarpy
Cytokinins	 <p>Kinetin Zeatin</p> <p>6-Benzyl aminopurine</p>	The main function is the promotion of cell division. They help the development of embryos during seed development, delaying breakdown of chlorophyll and degradation of proteins in ageing leaves
Abscisic acid	 <p>Abscisic acid</p>	It is a negative growth regulator. It accumulates in many seeds and helps in seed dormancy. It has an important role as potential antitranspirant by closing the stomata

a large gap remains in respect to development of food products, food supplements, food born chemical toxicity and phytochemicals using in silico tools.

4.1 Antioxidants

Modern lifestyles, such as smoking, pollution, consumption of junk food etc. have posed severe threats to mankind through continuous production and accumulation of the deadly free radicals within the human system (Valko et al. 2016). Free radical

formation occurs continuously in the cells as a consequence of both enzymatic and non-enzymatic reactions. Oxygen molecules are indispensable for performing various systemic functions. The ability of the antioxidants to control this free radical attack leads the medicinal chemist's primary attention to the design and modeling of efficient synthetic antioxidant molecules. Nature constitutes an abundant source of antioxidants. Fruits and vegetables serve as surplus sources of antioxidants. Vitamins and carotenoids exhibit the maximum antioxidant activity. Besides these, minerals like selenium and several phytochemicals like flavonoids, polyphenols, lycopene, lutein and lignans also exert free radical scavenging activity by varying mechanisms. A very few of such molecules with antioxidant property have been synthesized till date and even a fewer among them have contributed to be efficient drugs and food preservatives. The emerging concept is that dietary and endogenous antioxidants, endowed with different activities and characteristics, work synergistically contributing to the overall protective effect of plant foods. Thus, several attempts are now being made by the medicinal chemists to develop new active antioxidant molecules with improved activity profile.

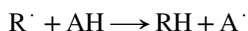
4.1.1 Antioxidants: The Free Radical Scavengers

The antioxidants are chemical agents that are capable of neutralizing these free radicals by inhibiting the oxidation process. According to US Food and Drug Administration (FDA), antioxidants are defined as "substances used to preserve food by retarding deterioration, rancidity or discoloration due to oxidation" (Halliwell and Gutteridge 1990). The antioxidants primarily function based on the following mechanisms:

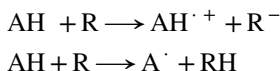
- Chain breaking reaction
- By reducing concentration of reactive oxygen species
- By scavenging initiating radicals
- By chelating transition metal catalyst
- Synergistic agents

4.1.2 Molecular Mechanism of Antioxidant Action

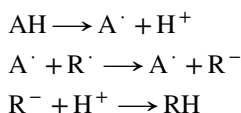
- **Hydrogen atom transfer (HAT):** The free radical removes a hydrogen atom from the antioxidant (AH) that itself becomes a radical (Wright et al. 2001).



- **Single-electron transfer followed by proton transfer (SET-PT):** The antioxidant can give an electron to the free radical becoming itself a radical cation. In the first step, electron transfer to a radical (R^\cdot) while in the second step, proton is transferred to the formed R^- anion (Musialik and Litwinienko 2005).



- **Sequential proton loss electron transfer (SPLET):** The reaction enthalpy of the SPLET first step corresponds to the proton affinity of the phenoxide anion (A^\cdot) (Fujio et al. 1981). In the second step, electron transfer from phenoxide anion to R. occurs and the phenoxyl radical is formed. The reaction enthalpy of this step denotes as electron transfer enthalpy.



4.1.3 Classification of Antioxidants

Various types of antioxidants are used for the benefit of human health classified as per their sources in Table 7 (Denisov and Afanasev 2005).

4.1.4 Screening Methods of Antioxidants

In vitro methods of screening are qualitative methods to determine antioxidant potency of a compound. However, IC_{50} values (concentration which can achieve 50% scavenging) or Trolox equivalents are used to quantify the activity (Kaur and Geetha 2006). Some of the antioxidant assay methods commonly used are briefly listed in Table 8.

4.1.5 Need of Synthetic Antioxidants

Considering the wide range of utility of the antioxidants, it may be hypothesized that the consumption of food rich in antioxidants and/or supplementation of antioxidants is important for leading a healthy livelihood. The antioxidants may not totally eradicate the occurrence of different free radical mediated diseases but may delay the progression of these diseases, and a proper diet rich in antioxidants may protect the human system from the attack of the fatal diseases. The huge importance

Table 7 Classification of different types of antioxidants

Types	Endogenous factors	Endogenous enzymes	Nutritional factors
Primary	Glutathione and other thiols	GSH reductase	Ascorbic acid (Vitamin C)
	Haem proteins	GSH transferases	Tocopherols (Vitamin E)
	Coenzymes Q	GSH peroxidases	B-carotene and retinoids
	Bilirubin	Superoxide dismutase	Selenium (essential dietary component of peroxidase)
	Urates	Catalase	Methionine or lipotropes for choline biosynthesis
Natural	Chemical group of antioxidant	Source	
	Vitamin A (retinol) and carotenoids	Carrots, squash, broccoli, sweet potatoes, tomatoes, cantaloupe, peaches and apricots	
	Vitamin C (ascorbic acid)	Citrus fruits, green peppers, broccoli, green leafy vegetables, strawberries and tomatoes	
	Vitamin E (tocopherol)	Nuts, seeds, whole grains, green leafy vegetables, vegetable oil and liver oil	
	Selenium	Fish, red meat, grains, eggs, chicken and garlic	
	Flavonoids/polyphenols	Red wine, purple grapes, pomegranate, soy, cranberries and tea	
	Lycopene	Tomato, pink grapefruit and watermelon	
	Lutein	Dark green vegetables such as kale, broccoli, kiwi, brussels sprout and spinach	
Lignan	Flax seed, oatmeal, barley and rye		
Secondary (synthetic)	Mechanism of action	Antioxidant	
	Break chains by reaction with the peroxy radicals	Phenols, naphthols, hydroquinones, aromatic amines	
	Break chains by reaction with alkyl radicals	Quinines, nitrones, iminoquinones	
	Hydroperoxide decomposing antioxidants	Sulfides, phosphites	
	Metal-deactivating antioxidants	Diamines, hydroxy acids	
	Cyclic chain termination	Aromatic amines, nitroxyl radicals	
	Inhibitors of combined action	Anthracene, methylenequinone	

of antioxidants has lead the researchers to search for synthetic antioxidants with improved activity and reduced toxicity. Thus, a great deal of recent research has been concentrated on the synthesis of antioxidants, which are as stable as the natural ones after loosing their electron (McCord 2004). The extent of damage caused by free radicals might be modified through three dietary intervention

Table 8 Commonly used methods for screening of antioxidant potency

In vitro methods of screening
Oxygen radical absorbance capacity (ORAC) method
1,1-diphenyl-2-picryl hydrazyl radical (DPPH) assay
Total radical-trapping antioxidant parameter (TRAP) method
Trolox equivalent antioxidant capacity (TEAC) method
Total oxyradical scavenging capacity (TOSC) method
Peroxyl radical scavenging method
Ferric reducing antioxidant power (FRAP) method
ABTS (2,2- α -azino-bis-(3-ethylbenzthiazoline-6-sulfonic acid) method
β carotene bleaching method
Gibb's Reagent (2,6-dichloroquinonechlorimine)
Nitroblue tetrazolium dye (NBT)
Hydrogen peroxide decomposition method
Assay based on extent of lipid peroxidation inhibition of antioxidants
Thiobarbituric acid assay
Iodometric method
Ferrous chloride and thiocyanate system
Ex vivo methods
DPPH assay
Singlet oxygen scavenging
In vivo methods
Lipid peroxidation
Nitric oxide scavenging bioassay

strategies: (a) caloric restriction and thus a depression in free radicals arising due to normal metabolism; (b) minimizing the intake of components that increase free radicals such as polyunsaturated fats; and (c) supplementation with one or more antioxidants. Thus, the ability of the antioxidants to control the free radical attack lead the medicinal chemist's primary attention to the design and modeling of efficient synthetic antioxidant molecules.

4.2 Food Supplements

The FDA regulates both finished food supplements and dietary ingredients under a different set of regulations than those covering "conventional" foods and drug products (FDA 2014). Under the Dietary Supplement Health and Education Act of 1994 (DSHEA):

- Manufacturers and distributors of dietary supplements and dietary ingredients are prohibited from marketing products that are adulterated or misbranded. That means that these firms are responsible for evaluating the safety and labeling of their products before marketing to ensure that they meet all the requirements of DSHEA and FDA regulations.
- FDA is responsible for taking action against any adulterated or misbranded dietary supplement product after it reaches the market.

A food and/or dietary supplement is projected to supply nutrients that may otherwise not be consumed in sufficient quantities. Supplements include vitamins, minerals, fiber, fatty acids, or amino acids, among other substances. U.S. authorities define dietary supplements as foods, while elsewhere they may be classified as drugs or other products. There are more than 50,000 dietary supplements available in present time according to US FDA. More than half of the U.S. adult populations (53–55%) consume dietary supplements with most common ones being multivitamins.

4.3 Flavoring Agents

Flavor is the sensory impression of a food or other substance, and is determined mainly by the chemical senses of taste and smell. The “trigeminal senses”, which detect chemical irritants in the mouth and throat, may also occasionally determine flavor. The flavor of the food, as such, can be altered with natural or artificial flavorants, which affect these senses (Smith et al. 2005). Of the three chemical senses, smell is the main determinant of a food item’s flavor. While the taste of food is limited to sweet, sour, bitter, salty, umami, and other basic tastes, the smells of a food are potentially limitless. A food’s flavor, therefore, can be easily altered by changing its smell while keeping its taste similar. As chemical structure is highly responsible for flavors, specific taste and smells, cheminformatics can play immense role for economical and effective development of flavoring agents for food industries.

Flavorants are focused on varying or enhancing the flavors of natural food products such as meats and vegetables, or creating flavor for food products that do not have the desired flavors. Most types of flavorants are focused on scent and taste. Few commercial products exist to stimulate the trigeminal senses, since these are sharp, astringent, and typically unpleasant flavors. As per the legal definition by the U.S. Code of Federal Regulations, a *natural flavorant* is: “the essential oil, oleoresin, essence or extractive, protein hydrolysate, distillate, or any product of roasting, heating or enzymolysis, which contains the flavoring constituents derived from a spice, fruit or fruit juice, vegetable or vegetable juice, edible yeast, herb, bark, bud, root, leaf or any other edible portions of a plant, meat, seafood, poultry, eggs, dairy products, or fermentation products thereof, whose primary function in food is flavoring rather than nutritional.”

Most artificial flavors are specific and often complex mixtures of singular naturally occurring flavor compounds combined together to either imitate or enhance a natural flavor. These mixtures are formulated by the flavorist to give a food product a unique flavor and to maintain flavor consistency between different product batches or after recipe changes. The National Association of Flavors and Food-Ingredient Systems (NAFFS) located in Neptune, New Jersey, USA is a broad-based trade association of manufacturers, processors and suppliers of fruits, flavors, syrups, stabilizers, emulsifiers, colors, sweeteners, cocoa and related food ingredients. The NAFFS' mission is to provide a forum for the exchange of technology and marketing information about food and flavor industries, and membership is open to all companies related to products and services to the food industry.

4.4 Phytochemicals

Foods of plant origin are an essential part of the diet; however, some plants contain substances that hold certain health risks also. Therefore, along with positive effects of phytochemicals, there are some phytochemicals which can induce adverse effects by different mechanisms (Kar and Roy 2012). Thus, some substances, for example, soy isoflavones, affect the endocrine system. Other substances are hepatotoxic (coumarin), neurotoxic (solanine), phototoxic (furocoumarins) or carcinogenic (estragole). Normal intake of phytochemicals as natural components of fruits, vegetables, herbs, and spices is regarded to be of low risk. Increased exposure, however, poses a potential problem, for example in cases of unbalanced diets or uptake of dietary supplements in isolated and concentrated form. For risk assessment, it is necessary to identify potentially harmful substances based on their chemical structure. Also, their dose-dependent effects on the organism have to be described. For this purpose, novel profiling techniques and advanced computational methods are increasingly used, with promising possibilities.

5 Food Safety and Regulatory Authorities

Food safety monitoring is a great challenge in present times due to adulteration of agrochemicals and food products. Therefore, the safety issue should be checked from the beginning of agriculture to a particular finished food product as safety can be affected at any time of food preparation. Various regulatory agencies are implementing as well as still preparing diverse white papers for proper monitoring of agrochemicals in food products along with providing best quality food to the world wide consumers (Kroes et al. 2004).

In the United States (US), major agencies responsible for food safety are:

- The United States Environmental Protection Agency (EPA) has set pesticide tolerances (maximum residue limits) in foods.
- The United States Department of Agriculture (USDA) and United States Food and Drug Administration (USFDA) are responsible for monitoring foods and animal feeds to ensure regulatory compliance with the tolerances. The US FDA is responsible for monitoring all foods with the exception of meat, poultry, and certain egg products. The USDA's Food Safety and Inspection Service (FSIS) also administers the national residue program and inspects sample of meat processed in slaughter plants for residue testing.
- The Federal Food, Drug, and Cosmetic Act (FFDCA) authorizes the US EPA to set tolerances with a safety standard in order to meet a "reasonable certainty of no harm".

The pesticide safety threshold limit applies to the food products from USA as well as products imported into the USA intended for human consumption. All the mentioned federal agencies must coordinate among themselves to accomplish their mission of ensuring food safety and generating plan on how to upgrade the US food safety system (<http://www.foodsafetyworkinggroup.gov/>).

From the beginning of 21st century, in European Union (EU), various regulations were formed to maintain the food safety through harmonizing the use of agrochemicals and adulteration of food commodities. Significant regulations are discussed below which are the benchmark in the process of food safety:

- The European Parliament and the Council adopted Regulation (EC)178/2002 (EC 2002) in the year 2002 stated down the overall principles and necessities of food law and procedures in matters of food safety harmonizing the free movement of food and feed in the EU. The Regulation ensures the protection of plants and plant products along with the aim of increasing agricultural production through plant protection. It not only gives the regulation of use of proper pesticides to protect crops before and after harvest against harmful organisms as well as to check the possible presence of pesticide residues in the treated products. Therefore, Maximum residue levels (MRLs) are set by the European Commission at the lowest achievable level consistent with good agricultural practices to protect consumers from exposure to unacceptable levels of pesticide residues in food and feed.

A legislative framework Regulation (EC) No 396/2005 of the European Parliament and of the Council (EC 2005) on pesticide residues applies in the EU from 1 September 2008. The Regulation attains the synchronization and generalization of pesticide MRLs ensuring better consumer protection throughout the EU. With the implication of new rules, MRLs undergo through a collective assessment of all form of products for all classes of consumers, including children. The decision-making is strictly science-based and a consumer intake assessment is conceded by the European Food Safety Authority before concluding on the safety of an MRL.

A dietary risk assessment is therefore a prerequisite for any MRL-setting. A major difficulty stems from the fact that only the toxicological properties of the active substance are normally directly investigated through the range of toxicological studies required according to Directive 91/414/EEC (EC 1991). For this reason, and within the framework of a collaboration agreement between the European Food Safety Authority (EFSA) and the European Commission's Joint Research Centre (JRC), a project was initiated to evaluate the possible contribution of computational methods and, in particular, QSAR analysis, to the evaluation of the toxicological relevance of metabolites and degradates of active substances of pesticides for dietary risk assessment. This project was one of three pesticide metabolism related projects sponsored by EFSA during 2009–2010. The other two addressed the possible use of Threshold of Toxicological Concern (TTC) considerations in assessing metabolite/degrade toxicity carried out by the UK Pesticides Safety Directorate (CRD 2010) and the impact of metabolism and degradation on pesticide toxicity performed by the Austrian Agency for Health and Food Safety (AGES 2010). In the framework of the EFSA funded project “Applicability of QSAR analysis to the evaluation of the toxicological relevance of metabolites and degradates of pesticide active substances for dietary risk assessment” (JRC 2010), the use of computational tools in the field of food safety was investigated.

In addition to the formalised approach of QSAR analysis, it is possible to estimate chemical properties and endpoints by using a less formalised approach, based on the grouping and comparison of chemicals. The grouping approach can be used, for example, to support the results of QSAR analysis or to generate estimated data assuming that, in general, similar compounds will exhibit similar biological activity (ECHA 2008). The use of computational tools in regulatory bodies and the food industry have many different applications and provide practical examples of how these methods can be applied and adapted for food safety purposes:

a. Use in US EPA OSCPP (Office of Chemical Safety and Pollution Prevention), Office of Pesticide Programs

According to the Organization for Economic Cooperation and Development (OECD) guidance, the US EPA uses the definition of a residue which describes whether to include pesticide metabolites in food safety assessments (OECD 2009). Regarding the determinations of hazardous residues of concern and tolerance expression, the Office of Pesticide Programs relies on the expert committee judgement of the Residue Of Concern Knowledge based Subcommittee (ROCKS). The ROCKS committee used QSAR tools routinely as the expert tool for predicting the hazards of pesticide and its metabolites of potential concern. They performed following process:

- (i) The US EPA has initiated long-term vision for the application of integrated testing and assessment tools for pesticides.
- (ii) Application of QSAR for hazard assessment for metabolites, environmental degradates and impurities.

- (iii) Application of QSAR in establishing “Threshold of Toxicological Concern”, non-animal eye irritation study and antimicrobial hazard
- (iv) Application of QSAR in ecological risk assessment

b. Use in US FDA CFSAN (Center for Food Safety and Applied Nutrition), OFAS (Office of Food Additive Safety)

The OFAS at the CFSAN of the US FDA states that QSAR analyses can be used to supplement genetic toxicity testing or explore endpoints like carcinogenicity, reproductive and developmental toxicity. Currently FDA CFSAN OFAS uses QSAR as a decision support tool in the safety evaluation of food products along with such as packaging materials and antimicrobials used in the food production (Arvidson et al. 2010).

6 Need of Application of In Silico Approaches in Food and Agrochemical Sciences

Phytochemicals are ubiquitous in dietary sources and can be found in many regulated food products as components of natural mixtures (e.g., flavouring agents, botanicals) and botanical extracts used as ingredients in dietary supplements and botanical drug products. Regrettably, a frequent setback with these substances is the lack of toxicology data which are obligatory as well as helpful for evaluating the safety of chronic human exposure. Chronic toxicity of a chemical is often pivotal evidence for regulatory decision-making on the safety of the product. Diverse range of toxicity must be addressed before introducing these phytochemicals into the open market. The carcinogenicity end point is among the most imperative chronic toxicities used to assess risk for human exposure to chemicals and in safety evaluations of regulated products. The regulatory guidance of United States (US) recommends the use of 2-year rodent carcinogenicity studies in two species and sexes to support the safety of US Food and Drug Administration (FDA) regulated products (FDA 2002). Although there is a great need for rodent carcinogenicity study data of chemicals, relatively few substances, especially phytochemicals, have been tested for carcinogenicity.

There are numerous reasons for the lack of diverse toxicity test data, including the excessive financial cost, intensive resources (experts and review), and obviously long period of time required to conduct the study according to standardized protocols such as those described in US FDA guidance documents (FDA 2002). Therefore, in silico studies can play immense role in the context of hazard and risk characterization of these substances in the interest of protecting public health (Jacobs 2005).

Phytochemicals are often encountered as “active” constituents in mixtures for which there are little or no toxicological data. Given their high potential for human exposure through dietary sources in conventional food and supplement products, there is a practical need for assessing their toxicity using efficient and reliable methods. Moreover, natural products show a vast structural diversity. Phytochemicals are a class of substances that present a data-poor situation in assessment of toxicity. Thus, there is a huge need for efficient prioritization of phytochemicals in testing and screening of chemical toxicity.

The use of *in silico* methods is now supported in the EU by the enacted REACH legislation in reaction to public desire to reduce the use of animals in testing (EU 2006). Moreover, these methods have been recommended by the US National Research Council (NRC 2007), and are considered to be useful in setting testing priorities (Bailey et al. 2005; NRC 2007). *In silico* models of rodent carcinogenicity using QSAR analyses of phytochemicals have been previously reported to be a predictive tool indicating some degree of promise for predicting naturally occurring carcinogens derived from plants (Arvidson et al. 2008).

At the US FDA Center for Drug Evaluation and Research (CDER), Office of Pharmaceutical Science (OPS), and the Center for Food Safety and Applied Nutrition (CFSAN), Office of Food Additive Safety (OFAS), Division of Food Contact Notifications (DFCN), the use of computational toxicology software is being employed to help support regulatory decision-making in the safety evaluation of human pharmaceuticals, their metabolites, and impurities, and indirect food additives (Arvidson et al. 2008; Mayer et al. 2008).

6.1 What is an In Silico Approach?

A great deal of recent research has been oriented towards the modeling and design of new molecules with the aim of discovery of potent molecules having improved response (therapeutic activity, agrochemical activity, or food) and less toxicity. *In silico* approaches (Fig. 1) play a crucial role in this protocol of rational new molecule discovery. The QSAR methodologies are the important computational tools which deal with the correlation between biological activity/toxicity of a molecule and its structural features (Helguera et al. 2008). In a QSAR study, the variations of biological activity within compounds of a congeneric series are correlated with changes in measured or computed features of the molecules referred to as descriptors. A QSAR model developed employing a series of molecules with a definite response helps in screening large databases of new molecules for identifying potential compounds with the specific response (Perkins et al. 2003). It thus reduces the huge expenditure of money and time for the preliminary experimental studies. Moreover, the REACH (Registration, Evaluation and Authorization of Chemicals) guidelines for animal safety restrict the extensive use of animals for

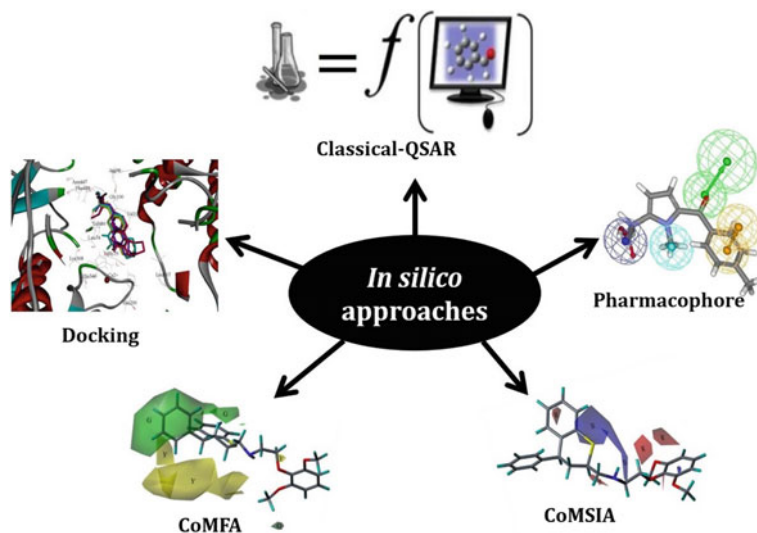


Fig. 1 Types of in silico approaches widely applied in new molecule design

initial screening of large databases. The QSAR technique thus provides an alternative pathway for design and development of new molecules with improved activity. The field of QSAR encompasses further studies related to quantitative structure-property relationship (QSPR) and quantitative structure-toxicity relationship (QSTR) (Kar et al. 2014). The QSPR study deals with the molecular features governing their physicochemical properties, while the QSTR technique determines the structural attributes of the molecules responsible for their toxicity profile. The pharmacophoric features and descriptors obtained from the developed QSAR models may also be utilized for virtual screening (Tikhonova et al. 2004) of large libraries of diverse compounds for a definite response parameter. A wide range of application of QSAR technique in the field of agriculture and food science is illustrated in Fig. 2.

A pharmacophore can be defined as the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response. The pharmacophore is generally defined by the following features, including H-bonding, hydrophobic, and electrostatic interaction sites, defined by atoms, ring centers, and virtual points. The pharmacophore does not represent a real molecule or a real association of functional groups, but a purely abstract concept that accounts for the common molecular interaction capacities of a group of compounds towards their target structure. Besides this, the identification of the prime features imparting improved activity to the molecules under a particular study facilitates the in silico design of new molecules with enhanced potency. Thus, a

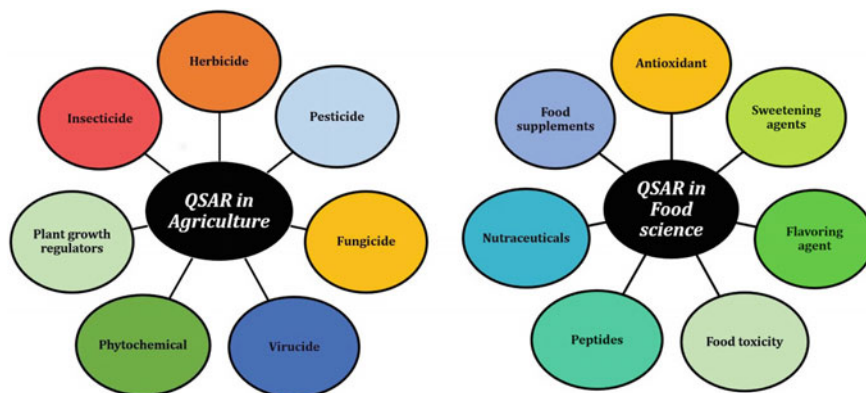


Fig. 2 Application of QSAR in the field of agriculture and food science

focused library may be developed by compiling the newly designed molecules with a specific response. QSAR analysis is based on the notion that biological activity (BA) depends on structure (C) and physicochemical properties (P) of the molecules. In case of QSTR, activity profile is replaced with the toxicity profile (Roy et al. 2015a, b).

$$\text{Biological activity/toxicity} = f(\text{Structure, Physicochemical properties})$$

There are various types of QSAR techniques evolved with the time, from 1D-QSAR to 7D-QSAR based on the descriptor dimensionality. Some popularly used methods are Molecular field analysis (MFA), Comparative molecular field analysis (CoMFA), Comparative molecular similarity indices (CoMSIA), multi-variate image analysis-quantitative structure-activity relationship (MIA-QSAR), Group based QSAR (GQSAR), Hologram QSAR (HQSAR,) etc. More details are discussed elsewhere (Roy et al. 2015a, b).

Another important in silico technique is docking study when the user has the clear idea about the selective receptor where the ligand molecule will bind to show the response in the body. Molecular docking is a study of how two or more molecular structures, for example drug and enzyme or receptor of protein, fit together (Roy et al. 2015a, b). The ability to bind large molecules, such as other proteins and nucleic acids to form supra-molecular complex play an important role in controlling biological activity. The behaviour of small molecules in the binding pockets of target proteins can be described by molecular docking. The method aims to identify correct poses of ligands in binding pocket of a protein and to predict the affinity between ligand and the protein. It can be classified as: (i) protein-small molecule docking, (ii) protein-nucleic acid docking and (iii) protein-protein docking (Roy et al. 2015a, b).

6.2 *QSAR in Regulatory Perspective*

Considering the occurrence of about 30 million chemicals (Benfenati 2012) serving varying purposes of industry, academia and household consumption, the assessment of their harmful impacts towards the living ecosystem seems to be an alarming issue. Along with the determination of specific toxicological impact of chemicals, specifically pharmaceuticals and agrochemicals, it is also necessary to determine their fate and impact on the environment. The aim should be to perform ecotoxicological assessment of the chemicals before they are released into the environment. Evidently it is impracticable to engage animal models for determining hazardous impact of such a huge number of chemicals. Hence, the concerned regulatory bodies of different countries across the globe promote the use of promising alternative techniques in achieving the goal. Computational *in silico* techniques such as QSAR modeling, read-across etc. serve as potential alternative techniques within the scope of predictive toxicology minimizing the time and cost of the research. Although complete replacement of animal testing is not feasible, reliable *in silico* techniques can help in the categorization of the chemicals and thereby guiding the usage of environmentally friendly chemical agents.

6.2.1 **The Acceptance of QSAR Modeling as an Alternative Strategy to Animal Testing**

The deployment of different animal models as testing objects in various chemical and pharmaceutical suffers from the ethical issues. Various international agencies on animal rights and ethics such as People for the Ethical Treatment of Animals (PETA), Fund for the Replacement of Animals in Medical Experiments (FRAME) etc. encourage the usage of alternative methods to minimize and/or eliminate harmful effects exerted to the animals. The QSAR modeling formalism involves the minimal use of animal experiments and accordingly complies with the guidelines proposed by various animal ethics bodies. The commonest of them is the ‘Three Rs (3Rs)’ formalism of Russell and Burch (1959) viz. ‘Reduction’, ‘Refinement’ and ‘Replacement’ which advocates the use of humanly methods for treating the animal. Table 9 shows the ideology of the 3Rs.

Presently, various international regulatory authorities support QSAR technique and propose its application in biological and risk assessment of chemicals. Some of the organizations promoting the application of QSAR modeling include European Commission’s European Centre for the Validation of Alternative Methods (ECVAM) (Eskes and Zuang 2005), the Council for International Organizations of Medical Sciences (CIOMS 1985), the REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals) regulations of the European Union (Hengstler et al. 2006), Office of Toxic Substances of the US Environmental Protection Agency (Auer et al. 1990), the Agency for Toxic Substances and Disease Registry (ATSDR) (El-Masri et al. 2002), the OECD (2014) etc.

Table 9 Principles of 3R

Method	Brief description
Replacement	Replacement refers to the use of non-sentient material in place of living conscious beings. That is the implementation of methodologies and techniques which replace the animal experiments on human ground. Materials with emaciated nervous and sensory systems are usually designated as the non-sentient objects namely microorganisms, higher plants, metazoan endoparasites etc. It encourages the use of experiments involving in vitro systems such as tissues, whole and part cells, in chimico methods i.e., the use of synthetic macromolecule as proxy toxicity targets, in silico techniques involving chemometric modeling, non-testing approach like read-across, and other models of microorganisms, established animal cell lines, immature form of vertebrates, invertebrates etc.
Reduction	Reduction refers to the minimization of the number of animals used in the study. This corresponds to all possibilities to reduce the number of animals used per experiments. Some of the approaches are use of fewer animals to reach the same goal, obtaining more information from each animal, are duction in the number of animals implemented in the original methodology, by the avoid of engaging additional animals etc. Improved experimental design and statistical analysis, sharing of data between research groups and organizations, use of imaging methods enabling longitudinal studies in same animal etc.
Refinement	Refinement refers to the minimization of animal suffering, pain and distress when engaged in studies. The objective of refinement is to enhance the well-being of animals. Now, it is evident that reduction of distress or suffering during biological experiment also resembles reduced variability in results and thereby also improving quality of the data. Examples might comprise use of suitable anaesthetics and analgesics, training of animals to avoid stress during procedures such as blood sampling, necessary housing of animals e.g., nesting option for mice to get desired specific behaviour

6.2.2 QSAR and the OECD

The OECD promotes the use of QSAR approach under the financial assistance of the European Union. With the aim of highlighting QSAR modeling as a reliable tool for the safety assessment of chemicals, the member countries of the OECD has developed a set of guidelines enabling its practical application in the regulatory context. The OECD (Q)SAR project comprising the QSAR toolbox (<http://www.oecd.org/chemicalsafety/risk-assessment/theoecdqsartoolbox.htm>), the principles for the validation of developed models, and guidance document aims to improve the application of QSAR modeling by governments and industry to facilitate the evaluation of chemical safety (OECD 2014). The OECD agreed for the following set of five principles to facilitate the regulatory use of QSAR modeling.

(i) *OECD Principle 1: A defined endpoint*

This principle dictates transparency to be maintained while choosing an endpoint data for modeling. The QSAR models are expected to be developed using homogeneous datasets containing single protocol generated response data. Another

Table 10 The commonly employed regulatory endpoints identified by OECD for predictive modeling analysis

Physicochemical properties	Environmental fate and toxicity		
	Ecological effects	Human health effects	Environmental fate
Melting point	Acute fish toxicity	Acute oral toxicity	Biodegradation
Vapour pressure	Acute Daphnid toxicity	Acute inhalation toxicity	Bioaccumulation
Aqueous solubility	Algae toxicity	Acute dermal toxicity	Hydrolysis
Boiling point	Long-term aquatic toxicity	Skin irritation	Atmospheric oxidation
K-octanol/water	Terrestrial effects	Eye irritation	
K-organic C/water		Skin sensitization	
		Repeated dose	
		Genotoxicity	
		Reproductive toxicity	
		Developmental toxicity	
		Carcinogenicity	
		Organ toxicity	

crucial point a QSAR modeler needs to consider is the similarity in mechanism/mode of action for all the chemicals used. Table 10 presents the commonly used endpoints as identified by the OECD guidelines.

(ii) *OECD Principle 2: An unambiguous algorithm*

This principle states an explicit methodology to be used while developing predictive QSAR models. This includes the formalisms implemented during data pre-treatment, dataset division and the selection of features. Hence, this rule focuses to bring transparency in model building rendering it not only reproducible to others, but also making it explanatory in achieving the endpoint estimates.

(iii) *Principle 3: A defined domain of applicability*

The third principle of OECD for development and validation of QSAR model portrays the importance of chemical/response domain of applicability. A QSAR model developed using a set of chemicals possesses a specific theoretical space and it is considered to provide reliable predictive result within that domain. Netzeva et al. (2005) has defined the applicability domain (AD) of QSAR models as follows: "The applicability domain of a (Q)SAR model is the response and chemical structure space in which the model makes predictions with a given reliability." Hence, the determination of the domain of applicability of a model using the training set molecules is necessary to check whether the prediction of test set molecules is trustworthy or not. The AD of a model depends on three major attributes (a) structural information, (b) physicochemical feature, and (c) response space.

(iv) *OECD Principle 4: Appropriate measures of goodness-of-fit, robustness and predictivity*

The fourth OECD principle gives information on the statistical judgment of stability and predictivity of a model. The judgment can be made by determining different metrics characterizing the (a) internal model performance by fitness and robustness measure using the training set, and (b) external predictivity using a test set. The objective is to determine a suitable balance between the extreme conditions namely overfitting and underfitting of the model based prediction.

(v) *OECD Principle 5: A mechanistic interpretation, if possible*

The fifth OECD principle attempts to draw attention on the diagnostic feature of the variables aiding a good mechanistic basis for the response being modeled. Definite information on the mechanism of action of chemicals towards a process can guide the design and development of only desired analogues. Now, from the statement it is evident that furnishing mechanistic information may not always be feasible, and the rule suggests the modeler to report if any such information is available facilitating future research on that endpoint.

A quick look at the OECD guidelines for development and validation of QSAR models is shown in Table 11, while Table 12 gives definition of the some of the terminologies used in QSAR modeling as per the OECD (2014) guidance document.

7 Successful Application of QSAR in Agriculture

7.1 QSAR Models of Herbicides

The activity of 69 monosubstituted sulfonylurea analogs (Fig. 3) as inhibitors of pure recombinant *Arabidopsis thaliana* Acetohydroxyacid synthase (AHAS) was studied to develop 3D-QSAR models using comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) by Wang et al. (2005). The studied research demonstrated the abilities of the 3D-QSAR techniques to explain the different affinities of herbicidal sulfonylureas for *A. thaliana* AHAS. To check the quality and predictability of the QSAR models, the authors used different training sets and found that CoMFA and CoMSIA gave similar correlations of inhibitory data. The CoMSIA analysis suggested steric, electrostatic, hydrophobic and H-bond acceptor features requirements for increased potency of this class of inhibitors. Mapping of the resulting fields on to the crystal structure of the yeast enzyme showed that the steric and hydrophobic fields were in good agreement with sulfonylurea-AHAS interaction geometry. The authors concluded that for high potency AHAS inhibition, the sole heterocyclic substituent should be small and hydrophobic while the *ortho* substituent on the aromatic ring

Table 11 The OECD guidelines for QSAR model development and validation at a glance

<p>Principle 1: A defined endpoint</p> <ul style="list-style-type: none"> ▪ Clear definition of the scientific purpose. ▪ The ability of the model in serving regulatory purpose. ▪ Important experimental conditions affecting the measurement. ▪ Unit of the endpoint.
<p>Principle 2: An unambiguous algorithm</p> <ul style="list-style-type: none"> ▪ For SAR: Explicit information on substructure and its substituents. ▪ For QSAR: Explicit information on the equation, algorithm, and the descriptors.
<p>Principle 3: A defined domain of applicability</p> <ul style="list-style-type: none"> ▪ For SAR <ul style="list-style-type: none"> ❖ Description of limits using substructure information. ❖ Rules depicting the impact on the molecular environment of the substructure. ▪ For QSAR <ul style="list-style-type: none"> ❖ Inclusion/ exclusion rule using the range of response and descriptors. ❖ Graphical presentation of training set descriptor values with respect to the response.
<p>Principle 4: Appropriate measures of goodness-of-fit, robustness and predictivity</p> <ul style="list-style-type: none"> ▪ Internal performance <ul style="list-style-type: none"> ❖ Training set information <ul style="list-style-type: none"> • Number of compounds • Chemical names • Structural formulae • Values of all descriptors • Values of all response variables ❖ Information on the raw data indicating processing method implemented (if any) ❖ Descriptors selection <ul style="list-style-type: none"> • Technique used for selecting initial descriptor pool • Initial number of considered descriptors • The feature selection technique used for using final descriptors • The final numbers of descriptors present in the model ❖ Employed statistical techniques <ul style="list-style-type: none"> • Specification of the method • Used software/ tool • Information on the independent application of the model ❖ Statistical metric showing goodness-of-fit using the training set ❖ Information on model cross-validation/ resampling ▪ External predictivity <ul style="list-style-type: none"> ❖ Information on the use of training set independent test set ❖ Information the external validation <ul style="list-style-type: none"> • Number of test set chemicals • Chemical names of all compounds • Structural formulae of all compounds • Values of all descriptors • Values of all response variables ❖ Suitable explanations on <ul style="list-style-type: none"> • The method used for selecting test set • Information on the size of the test set • Information on the chemical representativeness with respect to training set • Specifications of the employed statistical methods
<p>Principle 5: A mechanistic interpretation if possible</p> <ul style="list-style-type: none"> ▪ For SAR: Description of molecular events caused by the substructure leading to response ▪ For QSAR: Physicochemical interpretation of the descriptors with respect to a known mechanism ▪ An indication whether the proposed basis is derived using <i>a priori</i> or <i>a posteriori</i> observation

Table 12 Terminologies defined in the OECD guidance document (OECD 2014) in relation to QSAR paradigm

Terms	Definitions
Applicability domain (AD)	The AD of a QSAR model is the response and chemical structure space in which the model makes predictions with a given reliability
Classification	Classification is the assignment of chemicals to one of several existing classes based on a classification rule. A class or category is a distinct subspace of the whole measurement space. The classes are defined a priori by groups of objects in the training set. The objects of a class have one or more characteristics in common, indicated by the same value of a categorical variable. The classification method attempts to develop a classification rule using training set objects having known classes and aims to apply on test set objects bearing unknown classes
Cluster analysis	Cluster analysis is the grouping, or clustering, of large data sets on the basis of similarity criteria for appropriately scaled variables that represent the data of interest. Similarity criteria (distance based, associative, correlative, probabilistic) among the several clusters facilitate the recognition of patterns and reveal otherwise hidden structures in the data
Collinearity	Collinearity is a situation where there is a linear relationship between two or more of the independent variables in a regression model. In practical terms, this means there is some degree of redundancy or overlap the variables. Interpretation of the effects of the independent variables is difficult in this situation, and the standard error of their estimated effects may become very large
Congeneric series	A group of chemicals with one or more of the following: a common parent structure, same mechanism of action, and rate-limiting step
Cross-validation	Cross-validation refers to the use of one or more statistical techniques in which different proportions of chemicals are omitted from the training set (e.g. leave-one-out [LOO], leave-many-out [LMO]). The QSAR model is developed on the basis of the data for the remaining chemicals, and then used to make predictions for the chemicals that were omitted. This procedure is repeated a number of times, so that a number of statistics can be derived from the comparison of predicted data with the known data. Cross-validation techniques can be used to assess the robustness of the model (stability of model parameters), and to make estimates of predictivity
Degradation	Chemicals that are released in the environment are subject to different (biotic and abiotic) degradation processes: biodegradation by microorganisms, photolysis by light, hydrolysis by water, oxidation by different oxidants (for instance, in the atmosphere by hydroxyl and nitrate radicals or by ozone)
Dependent variable	A dependent variable (y) is a variable modelled by an equation in which one or more independent variables (x) are used as predictors of the dependent variable
Discriminant analysis	Discriminant analysis refers to a group of statistical techniques that can be used to find a set of descriptors to detect and rationalize (in terms of a predictive model) the separation between activity classes

(continued)

Table 12 (continued)

Terms	Definitions
Expert system	Any formalized system, not necessarily computer-based, which enables a user to obtain rational predictions about the properties or activities of chemicals. All <i>expert systems</i> for the prediction of chemical properties or activities are built upon experimental data representing one or more effects of chemicals in biological systems (the database), and/or rules derived from such data (the rule base)
External validation	External validation refers to a validation exercise in which the chemical structures elected for inclusion in the test set are different to those included in the training set, but which should be representative of the same chemical domain. The QSAR model developed by using the training set chemicals is then applied to the test set chemicals in order to verify the predictive ability of the model. In the ideal validation process, the results of external validation will be used to supplement the results obtained by internal validation
Internal validation	Internal validation refers to a validation exercise in which one or more statistical methods are applied to the training set of chemicals. Internal validation results in one or more measures of goodness-of-fit, robustness of model parameters, and estimates of predictivity
Molecular descriptor	A molecular descriptor is a structural or physicochemical property of a molecule, or part of a molecule, which characterizes a specific aspect of a molecule and is used as an independent variable in a QSAR
Molecular modeling	Molecular modeling refers to the investigation of molecular structures and properties by using computational chemistry and graphical visualization techniques to provide a plausible three-dimensional (3D) representation of a chemical. It can refer to the modelling of small organic molecules, macromolecules (e.g. proteins, DNA), crystals and inorganic structures. The 3D structure of the molecule is usually obtained by a process of geometry optimization. The geometry-optimized molecule provides the basis for calculating molecular properties
Narcosis	Narcosis is the non-specific suppression of physiological functions by chemicals which bind reversibly to membranes and proteins. The effect is brought about by non-reactive chemicals and is thought to result from an accumulation of the toxicant in cell membranes, diminishing their functionality. The narcotic effect is reversible, so that an organism will recover when the toxicant is removed
Pattern recognition	Pattern recognition is the identification of patterns in large data sets, using appropriate chemometric methods. Examples are exploratory methods like Principal Component Analysis (PCA), Factor Analysis, Cluster Analysis (CA), Artificial Neural Networks (ANN)
Parameter space	The parameter space of a model is a multi-dimensional space in which the axes are defined by the descriptors of the model
(Model) Performance	The performance of a QSAR model refers to its goodness-of-fit, robustness and predictive ability in relation to a defined applicability domain. Model performance is established by using the techniques of statistical validation

(continued)

Table 12 (continued)

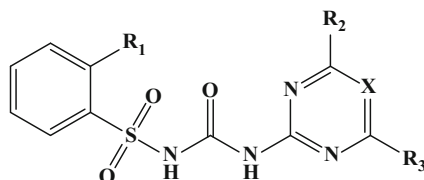
Terms	Definitions
Persistence	The term persistent is used to characterise chemicals that have long lifetimes in the environment. The persistence of a chemical depends on its kinetics or reactivity, as expressed by its rates of degradation
Predictivity	The predictivity (or predictive capacity/ability) of a model is a measure of its ability to make reliable predictions for chemical structures not included in the training set of the model
Reliability	Measures of the extent that a test method can be performed reproducibly within and between laboratories over time, when performed using the same protocol. It is assessed by calculating intra- and inter-laboratory reproducibility and intra-laboratory repeatability
Reliable (Q)SAR	A (Q)SAR that is considered to be “reliable” or “valid” for a particular purpose is a model that exhibits an adequate performance for the intended purpose. The criteria for determining whether the model performance is “adequate” will depend on the particular purpose and are highly context-dependent
Relevance	Description of relationship of the test to the effect of interest and whether it is meaningful and useful for a particular purpose. It is the extent to which the test correctly measures or predicts the biological effect of interest. Relevance incorporates consideration of the accuracy (concordance) of a test method
Structural alert	A structural alert is a molecular (sub)structure associated with the presence of a biological activity
Supervised learning	Supervised learning refers to the development of an algorithm (e.g. QSAR model) by a process that uses both the predictor and the response values, whereas in unsupervised learning, only the predictor values are used. Examples of supervised learning methods are (multiple) linear regression and discriminant analysis. Examples of unsupervised learning methods are different types of cluster analysis and principal components analysis (PCA)
Training set	A training set is a set of chemicals used to derive a QSAR. The data in a training set are typically organized in the form of a matrix of chemicals and their measured properties or effects in a consistent test method. A homogeneous training set is a set of chemicals which belong to a common chemical class, share a common chemical functionality, have a common skeleton, or common mechanism of action. A heterogeneous training set is a set of chemicals which belong to multiple chemical classes, or which do not share a common chemical functionality or common mechanism of action
Test set	A test set is sometimes called an “independent” or “external” test set (or validation set), and distinguished from “training set”. It is a set of chemicals, not present in the training set, selected for their use in assessing the predictive ability of a QSAR
Toxic endpoint	A toxic endpoint is a measure of the deleterious effect to an organism following exposure to a chemical. A large number of toxic endpoints are used in regulatory assessments of chemicals. These include lethality, generation of tumours (carcinogenicity), immunological responses, organ effects, development and fertility effects. In QSAR analysis, it is

(continued)

Table 12 (continued)

Terms	Definitions
	important to develop models for individual toxic endpoints, and different methods may be required for different endpoints
Validation	The process by which the reliability and relevance of a particular approach, method, process is established
Validated QSAR	A validated (Q)SAR is a model considered to be reliable for a particular purpose based on the results of the validation process in which the domain of application and the level of uncertainty required is defined
Valid (Q)SAR	A valid (Q)SAR is a model considered to be adequate for the intended purpose either because reliability has been demonstrated by historical use or by a validation process
(Q)SAR validity	The criteria for judging (Q)SAR validity are determined by specific regulatory constraints in member countries which include the number of chemicals, time required in the decision process and the level of uncertainty acceptable for the regulatory application

Fig. 3 Sulfonylurea herbicide scaffold employed by Wang et al. (2005)



should be small and polar. Compounds with more radical departures in structure from those of conventional sulfonylureas were generally ineffective inhibitors.

A series of pyrazolo[5,1-d][1,2,3,5]tetrazin-4(3H)one derivatives (Fig. 4) was designed, synthesized, and evaluated by Zhu et al. (2007) for their herbicidal activities towards *Brassica campestris* and *Echinochloa crus-galli* followed by QSAR studies employing physicochemical parameters (electronic, Verloop, or hydrophobic). The comprehensive study demonstrated that herbicidal activity against *B. campestris* was mainly affected by the molar refractivity (MR) for R^1 , Taft (Eso) for R^2 or R^6 , Verloop (Lm) for R^3 or R^5 , and electronic parameters (Hammett's constants) for R^4 . Again, herbicidal activity against *E. crus-galli* was mainly related with the substituents' hydrophobic parameter. When MR was about 0.95, the compound showed the highest herbicidal activity, and for the di-meta substituents (R^3 and R^5) at the benzene ring, their herbicidal activity was mainly affected by the substituent with higher Verloop's sterimol (L_m) parameter value. In the final conclusion the authors reported that these compounds showed greater herbicidal activity toward *B. campestris* than *E. crus-galli*.

A series of novel 2-cyanoacrylates (Fig. 5) containing different aromatic rings have been synthesized, characterized and tested for herbicidal activities against four weeds and inhibition of photosynthetic electron transport against isolated chloroplasts by Liu et al. (2008). Analysis of both in vivo and in vitro data showed that the

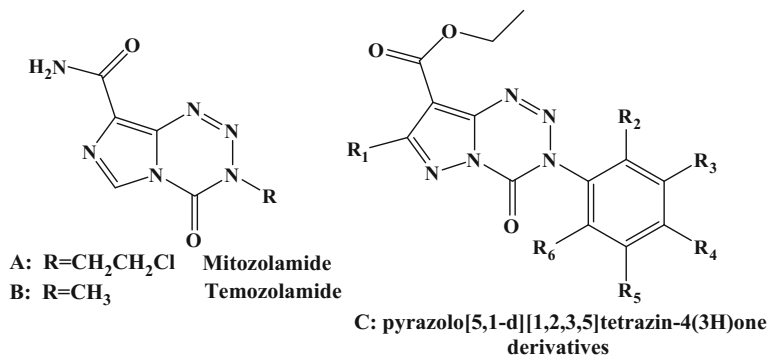
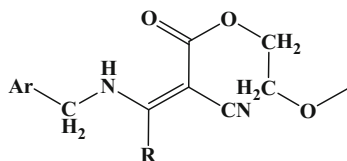


Fig. 4 Structure of Mitozolamide, Temozolamide and scaffold of pyrazolo[5,1-d][1,2,3,5]tetrazin-4(3H)one derivatives

Fig. 5 Skeleton of 2-cyanoacrylates



compounds containing benzene, pyridine, and thiazole moieties gave higher activities than those containing pyrimidine, pyridazine, furan, and tetrahydrofuran moieties which lead to further exploration through QSAR models on the basis of *in vitro* data. Therefore, the authors performed CoMFA analysis and the results showed that a bulky and electronegative group around the *para*-position of the aromatic rings would have the potential for higher activity, which offered important structural insights into designing highly active compounds prior to the next synthesis. The authors presumed that there was a significant electrostatic interaction between the aromatic ring and the possible receptor

A series of 68 sulfonylurea herbicides was modelled employing the MIA-QSAR approach by Bitencourt and Freitas (2008). The authors showed that the reported model seemed to exhibit advantages over traditional QSAR models correlated with physicochemical descriptors like log P (octanol/water partition coefficient) and MR (molar refractivity) through multiple linear regression (MLR) approach. Not only that the result of the presented study was compared with previous CoMFA model and found to be better predictive model in term of external validation. Therefore, the authors demonstrated that MIA-QSAR might be used as an alternative approach when existing methods do not work well and it had the capability for predicting new herbicide by taking a molecule that is a miscellany of substructures of known herbicides pertaining to two or more different congeneric classes having a minimum of similarity.

A predictive QSAR model was reported by Díaz and Delgado (2009) for pyrimidylsalicylate based herbicides for the prediction of 50% inhibition of the acetohydroxyacid synthase (AHAS) activity. The training set included 30 substituted O-(4,6-dimethoxypyrimidin-2-yl) salicylic acids and thio analogs, including 6-substituted(thio)-, 5- and 6-substituted salicylic acids covering pI_{50} range from about 3–8 U. The model was validated with an external set of 13 structures not included in the training set. Acceptable statistical results support the quality ($R^2 = 0.89$) and predictability ($R_{\text{Test}}^2 = 0.84$) of the developed QSAR model. The model involves only four descriptors: two geometric and two quantum chemical, accounting for the steric, electrostatic and hydrogen bonding interactions responsible for the binding of the herbicides to the enzyme. The result suggested that pyrimidylsalicylates and sulfonylureas have very similar binding sites which are in complete agreement with the literature.

Molecular docking-guided active conformation selection was used in a QSAR study for a series of 35 3H-pyrazolo[3,4-d][1,2,3]triazin-4-one derivatives as novel protoporphyrinogen oxidase (PPO) inhibitors with herbicidal activities by Lei et al. (2009). Molecular docking study was carried out to dock the inhibitors into the PPO active site and to obtain the rational active conformations. Based on the conformations generated from molecular docking, satisfactory predictive results were reported by a genetic algorithm-MLR (GA-MLR) model according to the internal and external validations (R^2 of 0.972 and 0.953 and an absolute average relative deviation AARD of 2.24% and 2.75% for the training set and test set, respectively). The four molecular descriptors contributed to the herbicidal activities of the studied compounds. Among them, the fragment C-033 (an atom centered fragment descriptor defined for each ring atom with three neighbours which represents the number of the R-CH...X fragment in a molecule with the meaning that a central carbon atom (C) on an aromatic ring has a carbon neighbor (R), a heteroatom neighbor (X) and the third hydrogen (H) neighbor outside the ring. “-” and “...” stand for aromatic and aromatic single bonds, respectively) played the essential role in the correlation between inhibitors and PPO, which can also explain the lower activities of few compounds. Moreover, a topological charge index descriptor and a R-GETAWAY descriptor indicated that the molecular conformation is greatly important in building the QSAR model. The outcome demonstrated that the molecular docking-guided active conformation selection strategy was rational and useful in the QSAR study of these PPO inhibitors and for the quantitative prediction of their herbicidal activities.

A QSAR study has been performed for a data set of 33 diphenyl ether (DPE) herbicides (Fig. 6) with their inhibition data on protoporphyrinogen oxidase (PPO) enzyme by Rouhollahi et al. (2010). PPO is the last common enzyme in the biosynthetic pathway to heme and chlorophyll. First, stepwise regression as a variable selection method was employed to develop a regression equation based on 26 training compounds, and predictive ability was tested on 7 compounds. Thereafter, two linear correlating tools, MLR and partial least squares (PLS) regression methods were used for final models. A multi-parametric equation

Fig. 6 Structure of Diphenyl Herbicides derivatives employed by Rouhollahi et al. (2010)

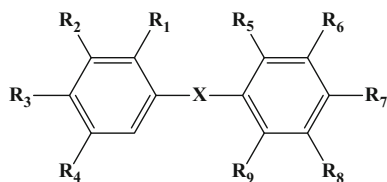
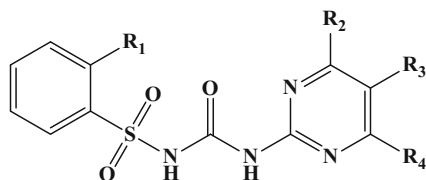


Fig. 7 Common scaffolds of sulfonylurea derivatives used by Roy and Paul (2010a)



containing four structural descriptors with good statistical impact was reported. Obtained results concluded that the *PCR* (ratio of multiple path counts to path counts), *T(Cl.Cl)* (sum of topological distances between Cl and Cl atoms), *RDF075m* (radial distribution function at 7.5 Å interatomic distances weighted by atomic mass) and *Mor03m* (3D-MoRSE—signal 03/weighted by atomic masses) can be used successfully for modeling biological activity of the studied compounds. Combination of 2D and 3D descriptors allowed differentiation between activities of enantiomers. The high correlation coefficients and low prediction errors obtained confirm good predictive ability of both MLR and PLS based models; resulting the squared regression coefficients were 0.95 and 0.94 respectively.

Docking and 3D-QSAR studies were carried out by Roy and Paul (2010a) for AHAS inhibitor sulfonylurea analogues (Fig. 7) having potential herbicidal activity. Docking studies suggested that the molecules bind within a pocket of the enzyme formed by important amino acid (Met351, Asp375, Arg377, Gly509, Met570 and Val571) residues. The inhibitors form hydrogen bonds with some of the amino acid residues to bind properly with the enzyme. But steric bumps have detrimental effect on the AHAS inhibition activity. The AHAS inhibitory activities of those compounds are very high which can form intramolecular or intermolecular hydrogen bonds. On the contrary, compounds, which formed steric bumps (either intermolecular or intramolecular), had lower AHAS inhibitory activity. Shape parameter showed that bulky substitution at R_1 position may enhance the AHAS inhibitory activity. The charged surface area descriptors recommended that the negative charge distributed over a large surface area may enhance the activity. The structural parameter (HBondacceptor) supports the charged surface area descriptors in that, for better activity, number of electronegative atoms present in a molecule should be high. The spatial descriptors proposed that for better activity the molecules should possess bulky substituent and small substitution at R_2 position and R_3 position respectively.

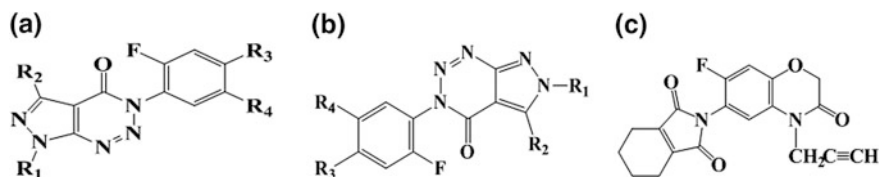


Fig. 8 Parent structures of 3H-pyrazolo [3,4-d][1,2,3] triazin-4-one derivatives and flumioxazin

Docking and 3D-QSAR studies have been performed by Roy and Paul (2010b) for protoporphyrinogen oxidase (PPO) inhibitor 3H-pyrazolo[3,4-d][1,2,3] triazin-4-one analogues (Fig. 8) which are potential herbicides to protect agricultural products from unwanted weeds. Docking studies suggested that the molecules bind with a hydrophobic pocket of the enzyme formed by some nonpolar amino acid (Ile168, Ile311, Ile412, Met365, Phe65 and Val164) residues. The co-enzyme FAD plays a major role in the receptor binding of the inhibitors. The inhibitors form hydrogen bonds to bind properly with the enzyme. But, steric bumps have detrimental effect on the PPO inhibition activity. Compounds which formed bumps either intermolecular or intramolecular way, had lower PPO inhibitory activity. The quantum chemical descriptor *LUMO* (energy of lowest unoccupied molecular orbital representing the electrophilicity of a molecule) suggested that, for better herbicidal activity the molecules should be highly electrophilic. But, another electronic descriptor *HOMO* (the energy of highest occupied molecular orbital representing nucleophilicity of a molecule) also showed positive contribution. So, there must be a balance between *HOMO* and *LUMO* energies, i.e., electrophilic and nucleophilic characters of the inhibitors. The charged surface area descriptors suggested that the positive charge distributed over a large surface area may enhance the activity. The spatial descriptors showed that for better activity the molecules should have symmetrical shape in all directions in a 3D space. Molecular field probes recommended that an increase in steric volume may enhance the herbicidal activity. Additionally, the position of the R_1 substituent may affect the PPO inhibition activity. Finally, authors concluded that instead of triazin-4-one ring system, tetrahydroisindole-1,3-dione ring structure may enhance the PPO inhibition activity.

To obtain insights into what/how properties and groups of the 12 phenylurea herbicides (PUHs) molecules affect the antigen-antibody (Fig. 9) interaction quantitatively, QSAR methodologies were applied including traditional 2D-QSAR and hologram QSAR (HQSAR) by Yuan et al. (2011). Both models showed high predictive abilities with cross-validated Q^2 values of 0.820 and 0.752, respectively. The results demonstrated that (1) the most important impact factor on PUH antibody recognition was the PUHs' hydrophobicity ($\log P$), which provides a quadratic correlation to the antibody recognition. Hapten carrier linking groups were less exposed to antibodies during immunization; thus, groups of the analytes in the same position were generally considered to be less contributive to antibody recognition during immunoassay. But the results of substructure-level analysis showed that

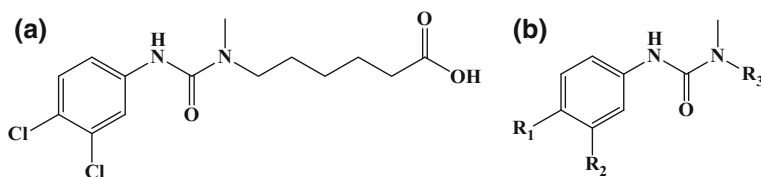


Fig. 9 Chemical structure of the hapten defines in A and common structure of PUHs is illustrated in B

these groups played an important role in the antigen-antibody interaction. (2) The frontier-orbital energy parameter E-LUMO was accountable for antibody recognition. (3) Although the R₃ group is less exposed to antibodies and considered to be less contributive, the hydrophobicity of the R₃ group has a great positive contribution to antibody recognition. To lengthen the R₃ group may increase the sensitivity of detection. (4) The R₁ and R₂ groups have negative contributions to antibody recognition; however they contribute positively or neutrally when they are with chlorine atoms on both groups that are same as the hapten.

Virtual screening analysis has been performed for 19 commercially available isatin analogues and 13 newly synthesized isatin derivatives as novel AHAS inhibitors for their herbicidal activity by Wang et al. (2011). The CoMFA contour models followed by density functional theory (DFT), natural bond orbital (NBO) and docking studies were prepared to understand the SAR for isatin derivatives. Combination of CoMFA and NBO approaches suggested that isatin moiety in all of the molecules makes a major contribution to the binding of *Arabidopsis thaliana* (AtAHAS). These were largely through π - π or hydrophobic interactions and most obvious in the HOMO maps. The frontier molecular orbital maps showed that the carbonyl positioned adjacent to the hydrazone nitrogen atom also played an important role in binding. Docked isatins to the binding site of AtAHAS demonstrated that the inhibitor fits neatly into the binding cavity and interact directly with AtAHAS. Docked interactions emphasized that hydrophobic contacts to the protein were through the indole moiety and the carbonyl carbon atom in the bridge. Docking binding mode was in good agreement with frontier molecular orbital from DFT calculations.

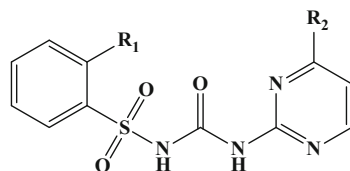
Mutation dependent Biomacromolecular QSAR (MB-QSAR) successfully applied by He et al. (2013) to quantitatively describe molecular mutational resistance to herbicide chlorimuron ethyl (CE) by AHAS which is one of the most important targets for herbicides, fungicides and antimicrobial compounds. The MB-QSAR models were constructed and validated based on the wild type *E. coli* AHAS II (EC 2.2.1.6) and its 85 mutants. The MB-QSAR models here gave accurate prediction of the pK_i app values for CE against AHAS mutants (MB-QSAR/CoMFA: $R^2 = 0.927$, $Q^2 = 0.631$, $R^2_{\text{pred}} = 0.684$; MB-QSAR/CoMSIA: $R^2 = 0.940$, $Q^2 = 0.540$, $R^2_{\text{pred}} = 0.690$). Interpretation of the 3D molecular interaction diagram gave detailed information about the structure resistance relationships for the mutated

AHAS. According to the authors, due to the limited available structural variables for the modification on small molecular ligands, 3D-QSAR of small molecular ligands may not give more comprehensive information on the intermolecular interaction than the MB-QSAR does. Therefore, MB-QSAR might provide further molecular information for designing the high potency inhibitors.

Most commonly, herbicides are ethyl to octyl esters of 2,4-dichlorophenoxy-acetic acids (2,4DAA), 2,4-dichlorophenoxy-propionic acids (2,4DPA) or 2,4-dichlorophenoxy-butyric acids (2,4DBA). Percutaneous penetration of esters of the 2,4D family was esterase-dependent both in rats and humans. The enzymatic constants for hydrolysis of each ester by skin esterases were determined in vitro using skin homogenates from both species. Therefore, SAR linking the evolution of the ex vivo percutaneous flux of esters and the 2,4D structure with enzymatic and/or physical parameters were examined by Beydon et al. (2014) to develop a good flux estimation model. The developed model suggested that although the percutaneous penetration of all of the esters were “esterase-dependent”, the decreasing linear relationship between percutaneous penetration and hydrophobicity defined by the logarithm for the octanol-water partition coefficient ($\log(k_{ow})$) was the most pertinent model for estimating the percutaneous absorption of esters for both species. The mean flux of the free acid production by the esterases of the skin was not the limiting factor for percutaneous penetration. The rate of hydrolysis of the esters in the skin decreases linearly with $\log(k_{ow})$, which would suggest that either the solubility of the esters in the zones of the skin that were rich in esterases or the accessibility to the active sites of the enzyme was the key factor. The structure-activity relationship resulting from this study makes it possible, in humans and in rats, to make a good estimate of the ex vivo percutaneous fluxes for all pure esters of this family of herbicides.

Bioactive compounds could form aggregates that influence the bio-interactive processes. Following the theory and based on π - π stacking models, quantitative aggregation-activity relationship (QAAR) studies were performed on a series of 24 sulfonylurea herbicides (Fig. 10) with good solubility by Xia et al. (2014). First, the authors generated π - π stacking aggregate models by the B97D/TZVP method. Then, QAAR investigations were performed on the descriptors calculated from the optimized aggregate/monomer structures. Four QAAR models were constructed, which indicated that the bioactivity may strongly depend on both the characters of the dimeric aggregates and the monomer. The best QSAR equation explored radius of gyration (r_{gyr} -a standard measure of overall structural change of macromolecules), \log of the aqueous solubility ($\log S$), \log of the octanol/water partition

Fig. 10 Common scaffolds of sulfonylurea derivatives used by Xia et al. (2014)



coefficient ($\log P_{(o/w)}$), surface globularity (v_{surf_G}) contributed to the herbicidal activity. The QAAR model revealed that low values of $rgyr$ and v_{surf_G} for the formation and dissociation of dimers, as well as high values of $\log S$ and low values of $\log P_{(o/w)}$ would lead to high bioactivity. The authors revealed that the QAAR approach was not only appropriate for poorly water-soluble insect growth regulators, but also for highly water-soluble sulfonylurea herbicide, and not only hydrogen bonds but also π - π interactions successfully introduced in QAAR investigations. The research work concluded that the QAAR approach based on dimer-aggregates can be applicable for the highly water-soluble sulfonylurea herbicides that can form π - π stacking interactions.

The herbicidal activities of a series of novel [1,2,4]triazolo[4,3-a]pyridine derivatives against monocotyledon weeds such as *Echinochloa crusgalli*, *Digitaria sanguinalis* and *Setaria faberii* and dicotyledon weeds such as *Amaranthus retroflexus*, *Eclipta prostrata* and *Brassica juncea* were evaluated by Liu et al. (2015), followed by a molecular modelling study employing the CoMFA method. A predictive CoMFA model was established with acceptable correlation coefficient R^2 (0.892) and the cross-validated coefficient Q^2 (0.61). The contributions of steric and electrostatic fields were 78.3% and 21.7% respectively. The 3D-QSAR provided meaningful clues as to the structural features of this family of herbicides that will be helpful in the design of more potent compounds in the future. The QSAR analysis indicated that the substituents on the benzene ring greatly affect the activity. The herbicidal activity with an electron-donating group is higher than that with an electron-withdrawing group. Compounds with an electron-donating group at the *para* position of the benzene ring exhibited significant herbicidal activity. In the halogen-substituted compounds, only fluorine substituted at the *para* or *meta* position of the benzene ring exhibited excellent herbicidal activity. However, it is noteworthy that the compounds with an alkyl group showed weaker herbicidal activity compared with that of aromatic-substituted compounds which showed moderate herbicidal activity. The effects on herbicidal activity can be placed in the order aromatic > heteroaromatic > alkyl.

The half-life ($t_{1/2}$) of 58 herbicides was modelled employing QSPR analysis by Samghani and Fatemi (2016). MLR and support vector machine (SVM) methods were used as feature mapping techniques for modelling and prediction the half-life after feature selection through stepwise-MLR from the large pool of descriptors. The statistical parameters R^2 and standard error for training set of SVM model were 0.96 and 0.087, respectively, and those were 0.93 and 0.092 respectively for the test set. The established SVM model was used for predicting the half-life of other herbicides that are located in the applicability domain of model determined via the leverage approach. The proposed models could identify and provide insight into what structural features were related to the half-life of these compounds. The result showed that the SVM model exhibits more reliable statistical and prediction performance than the MLR model. The good agreement between experimental results and the predicted values of half-life by using SVM indicated that the relationship

among selected molecular descriptors and herbicide's half-life is non-linear. The result emphasised that the process of degradation of herbicides in the environment is very complex and can be affected by the geometrical and topological aspects of molecules.

7.2 QSAR Models of Fungicides

A QSAR model was developed using a topological substructural molecular design (TOPS-MODE) approach to interpret the antifungal activity of 14 benzohydrazides which were synthesized and evaluated for their in vitro antifungal activity against the phytopathogenic fungus *Botrytis cinerea* by Reino et al. (2007). The model described 98.3% of the experimental variance with a standard deviation of 4.02. The performed study demonstrated that in addition to N',N'-dibenzyl and N-aminoisindoline substitution, the presence of a substituent in the *ortho* position of benzoic acid was critical to the antifungal activity. The QSAR study revealed that while N'-benzyl substitution seems to play an important role in the inhibition mechanism by enhancing the antifungal activity of these compounds, the aliphatic chain on the nitrogen of benzohydrazide does not seem to have a significant effect on their fungistatic activity, except when the aliphatic chain is sufficiently large for it to be considered appropriately hydrophobic. The authors concluded that the enhanced biological properties in that case might be caused by the necessary balance provided by hydrophobicity to cross the lipophobic or lipidic membranes of plants.

Three machine learning methods GA-MLR, least squares-SVM (LS-SVM), and project pursuit regression (PPR) were employed to develop the linear and nonlinear QSAR models by Du et al. (2008) for predicting the fungicidal activities of 100 thiazoline derivatives (Fig. 11) against rice blast caused by *M. grisea*. The authors used GA-MLR method to select the most appropriate molecular descriptors from a large set of descriptors followed by building of two models (LS-SVM and PPR) from selected ones. Both the linear and nonlinear models gave good prediction results, but the nonlinear models showed better prediction ability, which showed that the LS-SVM and PPR methods could simulate the relationship between the structural descriptors and fungicidal activities more accurately for this particular dataset. The study identified and provided significant insight into the structural

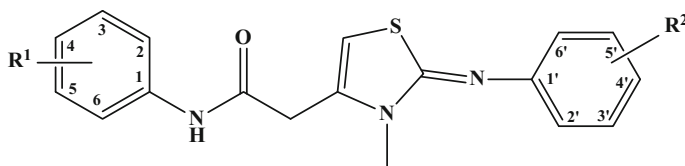


Fig. 11 Common scaffold of thiazoline derivative used by Du et al. (2008) and Song et al. (2008)

features related to the biological activity providing instruction for further design of thiazoline derivatives with higher inhibitory activity for the protection of rice blast disease.

Song et al. (2008) used fungicidal activities of 100 thiazoline derivatives (Fig. 11) as used by Du et al. (2008) to develop MLR and neural network (NN) models to study the substituent effects at *para* site of R¹ and at three sites (*ortho*, *meta*, or *para*) of R² aromatic rings in. Five descriptors including the non-overlap steric volume SV_{R²C²}, Connolly surface area SA_{R¹}, hydrophobicity $\sum \pi R^2$, and Hammett substituent constants (σ_{pR^1} and σ_{mR^1}) were identified as important factors of fungicidal activities. The authors reported the following conclusion from the study. For high fungicidal activities, substituents should have the small connolly surface area and electro-donation property at *para* site in R¹ aromatic ring. Fungicidal activities of thiazoline derivatives showed abilities and additive effects by the substituents as the small volume at *ortho* site, electron-withdrawing property at *meta* site, and high hydrophobicity in the R² aromatic ring. Again, the substituent effects in the R¹ aromatic ring were highly related to the fungicidal activities than those in R² aromatic ring. The correlations between the descriptors and the activities were improved by NN although the descriptors of optimum MLR model were used in the NN, which implies that the descriptors used in MLR model might have non-linear relationships with the response and these descriptors of thiazoline derivatives play a significant role in the fungicidal activities against *M. grisea*.

A series of 38 N-nitrourea derivatives (Fig. 12) were synthesized and their fungicidal activity were checked against *Rhizoctonia solani* by Cao et al. (2012) followed by construction of QSAR models using CoMFA and CoMSIA. Based on the experimental data, two best CoMFA and CoMSIA models with the cross-validated Q²_(LOO) values of 0.773 and 0.72 and correlation coefficient (R²) values of 0.959 and 0.936, respectively were obtained. The testing set of compounds gave a prediction (R²_{pred}) of 0.662 and 0.568 for CoMFA and CoMSIA models indicating that the two best models could be effectively used to predict the activity of new inhibitors and guide the further modification of these compounds by conforming to the following conclusion from the study. Small and electronegative groups at the 2-position (2-F), bulky and electronegative substituents at the 3-position (3-CF₃, 3-NO₂), bulky and electropositive groups at the 4, 5-region, and small and electronegative groups at the 6-position may increase the fungicidal activity.

Fig. 12 Basic structure of N-nitrourea derivative employed by Cao et al. (2012)

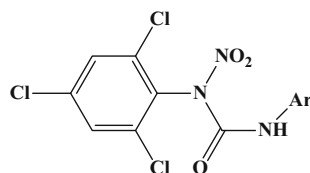


Fig. 13 Scaffold of (+)-DGA derivatives

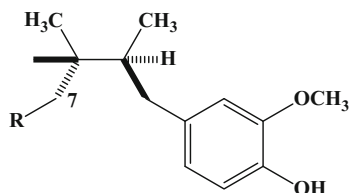
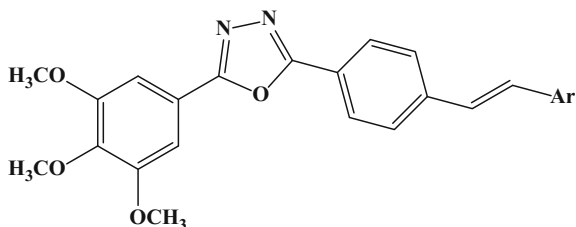


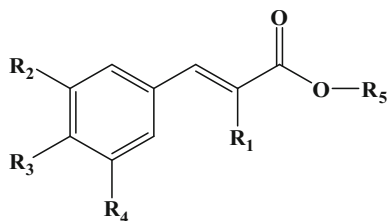
Fig. 14 Common structure of Stilbene derivatives containing the 1,3,4-oxadiazole moiety and trimethoxybenzene studied by He et al. (2015)



The relationship between antifungal activity against *Alternaria alternata* and structure on the 7-phenyl group of (+)-dihydroguaiaretic acid ((+)-DGA) was explained by investigating 38 synthesized (+)-DGA derivatives (Fig. 13) employing the Hansch-Fujita method by Hasebe et al. (2013). The model suggested that the small electron-withdrawing group at the meta-position of the 7-phenyl group is important for the higher antifungal activity. It was suggested that the smaller electron-withdrawing group at the 3-position caused higher activity. This was the first report of QSAR analysis of simple lignan, which was linked by only the β,β' -bond of two phenylpropanoid units. This result contributes to the design and synthesis of more active compounds based on the widely distributed simple natural lignan structure. For example, the whitening activity of 3-hydroxy-4-methoxyphenyl derivative, 3-hydroxy-4-ethoxyphenyl derivative, and 3-hydroxy-4-isopropoxyphenyl derivative against *A. alternata* Japanese pear pathotype was discovered.

He et al. (2015) investigated 22 novel stilbene derivatives (Fig. 14) containing the 1,3,4-oxadiazole moiety and trimethoxybenzene in 3D-QSAR analysis which were designed and synthesized previously. The 3D-QSAR approach proved to be an efficient one to implement structural modification by combination of several functional fragments and provided reliable clues for mechanistic study and designing of optimized stilbene derivatives in future. On the basis of the in vivo bioassay and 3D-QSAR analysis, the authors concluded that the variances (substituted positions, electronic properties, and steric effects) among substituents (Ar) on stilbenes showed a significant relationship with fungicidal activity against the three fungi (*Pseudoperonospora cubensis*, *Colletotrichum lagenarium* and *Septoria cucurbitacearum*). It is reported that the electron withdrawing substituents of the phenyl moiety seem to enhance the potency, which may be related to the fact that they tend to retard mechanisms of oxidative metabolism occurring on (or close to) the benzene ring. This assumption was consistent with the higher bioactivity of

Fig. 15 Basic skeleton of cinnamate derivative used by Saavedra et al. (2015)



the title compounds with electron withdrawing substituents at the *meta*-position as proposed by the 3D-QSAR analysis.

Fungicidal activities of a set of 27 active cinnamate derivatives (Fig. 15) were employed to investigate SAR by Saavedra et al. (2015). The authors explored constitutional, topological, geometrical and electronic molecular descriptors to predict the growth inhibition on *Pythium sp* and *Corticium rolfsii* fungi species; and the predicted values were in close agreement to the experimental values. From the developed QSAR model for the growth inhibition of *Pythium sp*, the authors revealed that that increased numerical values of RDF150 m (Radial Distribution Function—15.0/weighted by atomic masses) or decreased values of RDF020u (Radial Distribution Function—2.0/unweighted) or HATS2m (GATEWAY descriptor represent leverage-weighted autocorrelation of lag 2/weighted by mass) descriptors would lead to structures having a higher growth inhibitory activity on *Pythium sp*. The QSAR model for the growth inhibitory activity on *C. rolfsii* suggested that the increased numerical values of L3e or decreased values of BELv5 (Lowest eigenvalue n. 5 of Burden matrix/weighted by atomic van der Waals volumes) or RDF080u (Radial Distribution Function—8.0/unweighted) descriptors would lead to structures having a higher activity. Along with the development of QSAR models for two different species, the authors synthesized a set of 21 new structurally cinnamate compounds and predicted fungicidal activity. Finally, they have reported 3 and 2 cinnamates, expected to show higher activity for *Pythium sp* and *C. rolfsii*, respectively than the existing derivatives.

7.3 QSAR Models of Pesticides

A new modeling strategy based on the prioritization of fragments contribution to toxicity was developed and applied to 282 pesticides collected under the EU-funded project Demetra for predicting their toxicity toward the rainbow trout (*Oncorhynchus mykiss*) by Casalegno et al. (2006). Such toxicity models are important for the risk assessment of pesticides as well as development of better and greener pesticides. While there are other fragment based modeling routes, the authors exploited the possibility of top-prioritizing fragments' (TPFs) contributions to toxicity. On the assumption that one fragment might be mainly responsible for the molecular toxicity, they developed a three-stage modeling strategy to select the

most important moieties and to establish their priorities at a molecular level. Quantitative toxicity prediction yielded good results for the training set (R_{TR}^2 -0.85) and the test set (R_{TS}^2 -0.75). Analysis of the TPFs assigned during post assignment enabled to rationalize toxicity at the substructural level for the studied compounds.

Neuroblocking activity of imidacloprid analogs with various substituents at the 5-position of the pyridine ring measured and quantitatively analyzed using physicochemical substituent parameters by Nishimura et al. (2006). From the constructed QSAR model, they came to following conclusions. The greater the electron-releasing resonance effect, the higher is the pesticidal activity. The introduction of sizable and alkoxy substituents was unfavourable and substituents including halogens, alkoxy groups, alkyls and others into the 5-position of the pyridine ring of imidacloprid generally reduced neuroblocking activity. The reducing effect on blocking activity was well explained by the use of steric and electronic parameters. The introduction of alkoxy groups at this position was additionally unfavorable for activity. The neonicotinoids tested in this study probably bind first with nAChR, then cause blockage of the nervous system and kill the insects. A sequential scheme of this intoxication helped to understand the mode of action of neonicotinoid insecticides, and the QSAR results offered clues to vary the combination of 5th and 6th substituents on the pyridine ring or design other substituted heteroaromatic rings for new potential insecticides.

Bermúdez-Saldaña and Cronin (2006) investigated the development of QSARs for the toxicity to rainbow trout *Onchorhynchus mykiss* Walbaum of 75 organophosphorus and carbamate pesticides collected from a regulatory source, the US EPA database which is an openly available toxicological database. A large number of physicochemical and structural descriptors were calculated for the pesticides and QSAR models were developed using MLR and PLS tools. Because of the chemical heterogeneity of the dataset, relatively unsuccessful models were produced in term of predictive purposes. The authors reported that reducing the heterogeneity in the dataset by an MOA based approach gave better results than those based on chemical classes. The outcome supported that mechanistically based QSARs were likely to be more successful than those based on chemical classes. In the conclusion, they demonstrated that when variable selection was based on mechanistic approaches a highly relevant MLR model can be obtained.

3D-QSAR, docking, Local Binding Energy (LBE) and GRID methods were integrated together as combined tools for predicting toxicity and to explore the mechanisms of action on a set of 73 allelochemical-like pesticides particularly cyclic hydroxamic acids and lactams by Fratev et al. (2007). The 3D-QSAR model showed high predictive power, and the regression maps indicated the important toxic chemical substituents. The authors identified significant ligand-protein residue interactions and oxidation positions in the binding site by docking analysis employing CYP1A2 homology modelling. Computation of the binding energies of the compounds and the important substituents demonstrated quantitatively the substituent contributions in the metabolism and toxicity. The GRID examination identified the CYP1A2 binding pocket feature, and 3D-QSAR map was compared

to the GRID map. Interestingly outcome showed good overlaps confirming the important role of CYP1A2 in allelochemical like compounds toxicity.

Lu et al. (2007) explored the prediction of enzymatic activity of chloroperoxidase on metabolizing of selected Organophosphorus Pesticides (OPPs) by QSAR models. The authors employed quantum chemical descriptors computed with ab initio method at HF/6-31G(d) level and PLS analysis as the optimizing procedure for generating QSAR models. The correlation coefficient of the optimal model is 0.918, and the fitting results showed that it had high fitting precision and good predicting ability. The PLS assistant analysis indicated that the atomic charges of sulfur and phosphorus atoms in the S=P bond of an OPP molecule were important in governing the enzymatic activity and the molecular dipole moment also had some effect on the enzymatic activity. The authors found that OPPs with high absolute values of atomic charges on the sulfur and phosphorus atoms tended to be metabolized faster, whereas OPPs with stronger polarity tended to be metabolized slower by chloroperoxidase.

Nakagawa (2007) investigated classical QSAR for larvicidal and molting hormone activities and receptor-binding affinity of N,N'-dibenzoyl-N-t-butylhydrazines (DBH) using physicochemical parameters of the substituents. Considering QSAR of DBH for larvicidal and molting hormone activities, the analysis showed that larvicidal activity against larvae of *C. suppressalis* and *S. exigua* was enhanced with molecular hydrophobicity, but that the introduction of bulky substituents into any position on the A- and B-ring moieties, except for the *ortho* position of the A-ring, decreases the activity. Introduction of electron-withdrawing groups at the *ortho* position of the A-ring was favorable to the larvicidal activity to *C. suppressalis*. For larvicidal activity against *S. exigua*, the QSAR was similar to that for *C. suppressalis*, but, for that against *L. decemlineata*, it was somewhat different and a position specific optimum hydrophobicity of substituents was significant to govern the variations of the activity. The larvicidal activity of these compounds was lower than ecdysone activity which suggested that the metabolic detoxication of alkanoyl groups was significant under assay conditions. The QSAR study in the binding of DBH to the ecdysone receptor revealed the effects of the substituents at the *para*-position of B-ring of DBH on the binding affinity to Sf-9 cells quantitatively analyzed with the classical QSAR method. This analysis identified that the binding was enhanced with the hydrophobicity and electron-donating property of substituents, but decreased with increasing the width of substituents.

Ecotoxicological data based on the US EPA dataset consisting of 125 aromatic pesticides with aquatic toxicity towards trout was investigated using a QSAR analysis by Slavov et al. (2008). Additionally, the authors utilised an external test set of 37 compounds for validation purpose. Along with the standard 2D-QSAR analysis, the authors performed a CoMFA analysis considering the electrostatic and steric properties of the molecules. The CoMFA analysis helped the recognition of the steric interactions as playing an important role for aquatic toxicity. The best multilinear QSAR equation obtained with three variables: CODE_MID, molecular weight and heat of formation, all having positive regression coefficients. Analyzing

the outcome of the QSAR study, the following conclusions were made. The pesticides characterized by larger heats of formation were more stable and thus the probability to reach the target site unchanged was higher. Since no charge distribution related descriptors were involved into the model, they stated that the electrostatic interactions are of much lower importance for the aquatic toxicity than the steric interactions. This conclusion was fully supported by the outcome from the CoMFA study which stated that the steric interactions play a much more important role for the aquatic toxicity than the electrostatic interactions. The visual examination of the steric interactions map showed that the presence of bulky substituents around positions 3 and 4 of the aromatic ring and near the heteroatoms of the side chain will lead to an increased toxicity effect.

Wang et al. (2009) applied a combinatorial approach to a dataset of 1600 compounds with known aquatic toxicity ratings. The dataset consists of compounds spanning of five classes, i.e., the nontoxic, slightly, moderately, highly and very highly toxic pesticides, in which about 75% molecules selected randomly as a training set and the remaining one considered as external test set. Wang et al. (2009) established a series of classification based QSAR models of these pesticides. By an analysis of those statistically significant descriptors implicated in these SAR models, well-known theoretical descriptors such as the molecular weight, molecular connectivity indices, H-bond donor/acceptor electrotopological parameters were found dominating the models could be associated with a molecule's passive diffusion and binding affinity to cell membrane as well as the targeted proteins of organisms thus leading to its toxicity. These models can be used for estimating the aquatic toxicity rating of pesticide candidates at the early stages of pesticide discovery projects as well as for exploring their intrinsic toxic mechanisms.

Ruark et al. (2013) addressed the QSAR model to predict pentavalent organophosphate oxon human acetylcholinesterase bimolecular rate constants of a database consisting of 278 3D structures and their bimolecular rates. The database was quite diverse, spanning 7 log units of activity. The authors calculated 675 molecular descriptors employing AMPAC 8.0 and CODESSA 2.7.10. Orthogonal projection to latent structures regression, bootstrap leave-random-many-out cross-validation and y-randomization were used to develop an externally validated consensus QSAR model. The result showed that the HOMO–LUMO energy gap contributed most significantly to the binding affinity. A mean training R^2 of 0.80, a mean test set R^2 of 0.76 and a consensus external test set R^2 of 0.66 suggested robustness and predictability of the developed QSAR model. The outcome of this QSAR model can be used in physiologically based pharmacokinetic/pharmacodynamic models of organophosphate toxicity to determine the rate of acetylcholinesterase inhibition. This work fulfilled one of these data gaps through QSAR prediction of OP-AChE bimolecular rate constants which were the initial driver toward AChE inhibition. Ruark et al. concluded that once the active site of AChE is inhibited, it can be further modulated by the OP oxon aging and regeneration processes.

Tree-based multispecies QSAR models were constructed for predicting the avian toxicity of pesticides using a set of nine descriptors derived directly from the

chemical structures and following the OECD guidelines by Basant et al. (2015). The Bobwhite quail toxicity data was used to construct the single decision tree (SDT), decision tree forest (DTF), and decision tree boost (DTB) regression QSAR models and externally validated using the toxicity data in four other test species like Mallard duck, Ring-necked pheasant, Japanese quail, House sparrow. Intercorrelation analysis and PCA methods provided information on the association of the molecular descriptors related to molecular weight (MW) and topology. The S36 and MW were the most influential descriptors identified by DTF and DTB models. The DTF and DTB performed better than the SDT model and yielded a correlation (R^2) of 0.945 and 0.966 between the measured and predicted toxicity values in test data array. The results suggested for the appropriateness of the developed QSAR models to reliably predict the toxicity of pesticides in multiple avian test species and they can be useful tools in screening the new chemical pesticides for regulatory purposes. Substructural alerts were identified directly from mechanistic knowledge which was important to predict toxicity. In all five toxicity data sets, carboxamide (C(=O)N), carbonyl (C=O), aminocarbonyl (NC=O), aromatic amine (CN), aromatic halides (Ph-X), halide (X), nitro (N=O), sulfide (CS), phosphate thioate (P=O), oxophosphorus (P=O), aromatic alkane (Ph-C1), aromatic benzene, aliphatic ether (COC), and alkane linear (C2) were the common major substructures responsible for the avian toxicity.

Hamadache et al. (2016) developed a QSAR model to predict the oral acute toxicity of pesticides to rats employing 258 pesticides with an additional external set of 71 pesticides. Oral acute toxicity on rat was modeled based on the multi-layer perceptron-ANN (MLP-ANN) with descriptors calculated by Dragon software and selected by a step-wise MLR method. The seventeen selected descriptors showed that the electronic properties and the specific structural attributes of the molecule played the main role in the toxicity of the pesticides. The built MLP-ANN model assessed comprehensively based on internal and external validations parameters. Based on the comparison with previous models, the proposed QSAR model achieved better results and provided 98.6% predictions that belong to the applicability domain. The authors suggested that the model can be used to predict the acute oral toxicity of pesticides, particularly for those that have not been tested as well as new pesticides and thus help reduce the number of animals used for experimental purposes.

7.4 QSAR Models of Insecticides

The pharmacophore of the 23 variants of neonicotinoid insecticide was examined by Kagabu et al. (2008). Analysing the outcome of the *in silico* study, the authors concluded that most of the variations of the pharmacophore structure bearing NNO_2 , CHNO_2 , or NCN in the neonicotinoids afforded insecticidal activity against American cockroaches at nanomolar concentrations under synergistic conditions. The neuroblocking potency was proportional to the Mulliken charge on the nitro

oxygen atom or the cyano nitrogen atom, and was related to the log P value with the optimal value of 1.19. The potency of NCN compounds was lower by a factor of 2.03 in log units than the corresponding nitro compounds in the neuroblocking activity. Theories at various levels and experiments using several known molecules have predicted or speculated the crucial involvement of the presented structural fragments in the activity of neonicotinoids.

Li et al. (2008) successfully constructed an effective pharmacophore model (RMS = 0.634, Correl = 0.893, Weight = 1.463, Config = 11.940) based on a series of nAChR (nicotinic acetylcholine receptors) agonists, which consists of a hydrogen bonding acceptor, a hydrogen-bond donor, a hydrophobic aliphatic and a hydrophobic aromatic centre. They had designed a series of heterocyclic compounds by this pharmacophore model followed by synthesis of some of them. The developed pharmacophore model provided useful information for developing novel insecticides targeting at the nAChR in the near future.

Employing podophyllotoxin as a phytoinsecticidal lead compound, 15 novel aromatic esters of 4'-demethyl-4-deoxypodophyllotoxin (Fig. 16) were semisynthesized and preliminarily tested for their insecticidal activity against the pre-third-instar larvae of *Mythimna separata* Walker in vivo for the first time by Xu et al. (2009). Followed by synthesis, the authors performed QSAR studies of all 15 compounds and reported that the relative number of benzene rings and final heat of formation were very important properties to their insecticidal activity. A negative coefficient before the relative number of benzene rings indicated that increase of this value led to a decrease in the mortality rate of the compounds due to the reduction in the solubility of the compound. The final heat of formation was proportionally interrelated with the Gibbs free energy.

QSAR modeling was carried out for ovicidal activity of 2, 4-diphenyl-1, 3-oxazoline analogs (Fig. 17) against two-spotted spider mite *Tetranychus urticae* by Roy and Paul (2009). Models obtained by using 2D parameters revealed that the chain length of the substituent at *para* position of the 4-phenyl ring was a critical factor. Initially the ovicidal activity was enhanced as the substituent chain length

Fig. 16 Chemical structure of podophyllotoxin

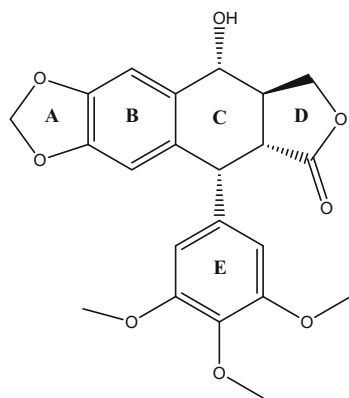
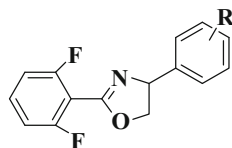


Fig. 17 Parent structure of 2-(2',6'-difluorophenyl)-4-phenyl-1,3-oxazoline derivative



increases, but after a certain limit the activity reduces though the chain length increases. This implied that the lipophilicity of the substituents should be optimum. Electrotological state indices of specific atoms (S_{17}, S_{16}) and substituent hydrophobicity parameter suggested that the presence of long chain *para* substituents containing electronegative atom directly attached to the 4-phenyl ring or at its close vicinity may increase the ovicidal activity. The value of substituent hydrophobicity constant at the *para* position should be within 1.98 to 2.90 for the optimal activity. Lipophilicity of the whole molecule also plays a dominant role. Models generated from 3D descriptors recommended that the shape of the substituents should be optimum and the lipophilic substituents having electronegative atoms with distributed positive charge over the surface may enhance the ovicidal activity. The model obtained from molecular field analysis suggested that bulky substituents with optimally distributed charge may increase the ovicidal activity.

Liu et al. (2010) explored 3D-QSAR and the pharmacophore models on 38 anthranilic diamides, potent activators of the anthranilic diamide ryanodine receptor (RyR) employing CoMFA, CoMSIA and distance comparison technique (DISCOtech) approaches. Computed CoMFA and CoMSIA models yielded acceptable cross-validated (q^2) values of 0.785 and 0.788 and non-cross-validated (r^2) values of 0.958 and 0.981, respectively. The obtained DISCOtech pharmacophore model indicated that hydrophobic interaction and hydrogen bonds had important roles in the interactions between activators and RyRs, which was consistent with CoMSIA results. The information obtained from CoMFA, CoMSIA and DISCOtech models enabled interpretation of the SAR of anthranilic diamides. Based on the constructed models, some vital features for the interaction of anthranilic diamides with RyRs were identified by the authors, which were helpful in designing more potent RyR activators.

Wei et al. (2011) isolated a series of 43 natural β -dihydroagarofuran sesquiterpene polyesters from *Celastrus angulatus* and *Euonymus japonicus* and evaluated their insecticidal or narcotic activities against the fourth-instar larvae of *Mythimna separata* followed by 3D-QSAR study employing CoMFA and CoMSIA analyses. The observed variances in the insecticidal and narcotic activities were explained from the optimal CoMFA and CoMSIA models. For the narcotic model, the electronic field explained as the most influential among all three fields, contributing 52.2% to the optimal QSAR model. A similar result also observed for the insecticidal model where the electronic field provided the highest contribution (38.8%) to the best model. A hybrid effect of the electrostatic (38.8%) and hydrophobic (40.2%) interactions governed the insecticidal activities of the molecules which indicated that the electronic interaction played the most important role in

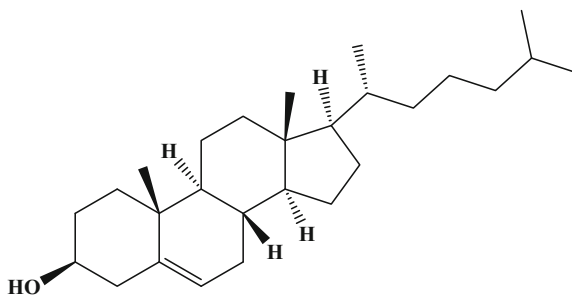
determining the biological activities of these molecules. The outcome from both the experimental and the theoretical investigations proved valuable for the design of novel β -dihydroagarofuran sesquiterpene polyesters with enhanced activities.

A multi-target QSAR (mt-QSAR) discriminant model was developed by Speck-Planche et al. (2012) for the search, design and prediction of insecticides acting through five different MoA using a large heterogeneous database consisting of 657 molecules. According to the authors, the explored model could predict insecticidal activity of compounds in more general situations than classical QSAR models which have as principal limitation the assessment of biological activity against only one type of MoA or biological receptor. The presented QSAR discriminant model classified correctly more than 90% of insecticides and inactive compounds in both, training and prediction series. The most striking part was that the model permitted the automatic and efficient extraction of fragments responsible of insecticidal activity against several mechanisms of action and new molecular entities were also suggested as possible multi-target insecticides according to the posteriori probabilities.

A series of 33 isoxazoline and oxime derivatives of podophyllotoxin modified in the C and D rings (Fig. 16) were synthesized and characterized by diverse analytical methods followed by their insecticidal activity was evaluated against the pre-third-instar larvae of northern armyworm, *Mythimna separata* (Walker) in vivo by Wang et al. (2012). To understand the responsible structural fragments and properties for insecticidal activity, the authors employed GA-MLR calculation performed by the MOBY DIGS package. QSAR studies demonstrated that the insecticidal activity of these compounds was mainly influenced by factors like electronic distribution and steric factors. The developed model attained the standard deviation error in prediction (SDEP) of 0.0592, the R^2 of 0.861, and the Q_{LOO}^2 of 0.797. Five descriptors were evolved from the model as the best possible features important for insecticidal activity as follows: 2D autocorrelation descriptor (GATS4e), edge adjacency indice (EEig06x), RDF descriptor (RDF080v), 3D MoRSE descriptor (Mor09v) and atom-centered fragment (H-052) descriptor. According to the standardized coefficient of the descriptors, EEig06x was identified as the most significant one calculated from the edge adjacency matrix of a molecule. The second important parameter GATS4e reflected the information on molecular dimension and Sanderson electronegativity. H-052 is defined as the number of specific atom types in a molecule and calculated by knowing the molecular composition and atom connectivity. All three descriptors were positively correlated to the activity. RDF080v interpreted as the probability distribution of finding an atom in a spherical volume of radius. The descriptor Mor09v was based on the idea of obtaining information from the 3D atomic coordinates by the transform and weighted by atomic van der Waal volumes.

Four series of novel cholesterol-based (Fig. 18) hydrazone derivatives were synthesized and their insecticidal activity was tested against the pre-third-instar larvae of oriental armyworm, *Mythimna separata* (Walker) in vivo by Yang et al. (2013) followed by QSAR study employing GA-MLR by the MOBYDIGS

Fig. 18 The chemical structure of cholesterol



software. Interestingly, the synthesized derivatives showed better insecticidal activity than their precursor cholesterol. The QSAR model demonstrated that six descriptors such as RDF085v (radial distribution function descriptor representing the molecular conformation in 3D with a series of weighting schema, such as weighted by atomic masses, atomic van der Waals volumes, atomic Sanderson electronegativities and atomic polarizabilities), Mor06u (3D-MoRSE-signal 06/unweighted parameter belongs to 3D-MoRSE descriptor), Mor11u (3D-MoRSE descriptor representing the 3D-MoRSE signal 23/unweighted), Dv (a WHIM descriptor, which represents D total accessibility index/weighted by atomic van der Waals volumes), HATS0v (a GETAWAY descriptor, which represents leverage-weighted autocorrelation of lag 0/weighted by atomic van der Waals volumes) and H-046 (an atom-centred fragment descriptor, which indicates the H attached to CO (sp³) no X attached to next C) were likely to influence the insecticidal activity of these compounds. Among them, two important ones were Mor06u and RDF085v. Mor06u provides 3D information from the 3D coordinates by using the same transform as in electron diffraction and RDF085v is related to van der Waals volumes of molecule.

Three novel series of N3-substituted imidacloprid derivatives were designed, synthesized, and characterized by NMR spectroscopy, mass spectrometry, elemental analysis, and single-crystal X-ray diffraction analysis by Wang et al. (2014). Thereafter, the insecticidal activities against *Aphis craccivora* were evaluated and QSAR analysis was performed by the authors. The QSAR results indicated that the size, electron density, and distribution of the substituents at the N3 position were critical to the derivatives' activity. Furthermore, the molecular docking analysis indicated that imidacloprid and active synthesized compounds form the similar hydrophobic and van der Waals interactions with Trp53, Met114, Trp143, Tyr185, and Tyr192 in the binding pocket. Compared with imidacloprid, more favorable van der Waals and hydrophobic interactions were identified from the methylbenzene group and the aromatic ring of chloro-phenyl group of most active compound. Stronger hydrogen bonding interactions from the nitro and sulfonyl group with the binding pocket contributed to higher insecticidal activity. The analysed results promoted the basic understanding about interaction mechanism of these derivatives

and nicotinic acetylcholine receptors (nAChR); and provided the useful information for further structural modification.

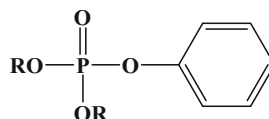
The bioactive conformations of anthranilic diamide insecticides were disclosed from a series of low energy conformations by Jiang et al. (2015) from the combined approach of DFT-based potential energy surface scanning and 3D-QSAR analysis like CoMFA and CoMSIA. From the thorough analysis, the authors concluded that an intramolecular N-H...O H-bond was found important for maintaining the bioactive conformation of chlorantraniliprole. For the bioactive conformation, if the phenyl was regarded as the paper plane and the methyl group connected with phenyl was pointing to the down direction, the O and H atoms forming the N-H...O H-bond were out the phenyl plane, and the pyridine was also pointing to the down direction with the substituted Cl inside the phenyl plane. In addition, the outcome also supported that DFT-based potential energy surface scanning combined with CoMFA analysis was a good approach for exploring the bioactive conformation of a compound, when the target structure was unknown.

A QSAR study was performed by Loso et al. (2016) using the sulfoximine insecticides to reveal the importance of a 3-pyridyl ring and a methyl substituent on the methylene bridge linking the pyridine and the sulfoximine moiety to achieve strong *Myzus persicae* activity. The SlogP driven regression model helped to explain the highly optimized pyridine substitution pattern for sulfoxaflor. The developed model was highly predictive one for an external set of 18 sulfoximines including sulfoxaflor. The model was consistent with and helped in explaining the highly optimized pyridine substitution pattern for sulfoxaflor.

Seifert (2016) determined the structural requirements of organophosphorus insecticides (OPI) for reducing chicken embryo nicotinamide adenine dinucleotide (NAD⁺) content in OPI-induced teratogenesis through 3D-QSAR analysis. The COMFA approach revealed the electrostatic and steric fields as good predictors of OPI structural requirements to reduce NAD⁺ content in chicken embryos. The dominant electrostatic interactions were localized at nitrogen-1, nitrogen-3, nitrogen of 2-amino substituent of the pyrimidinyl of pyrimidinyl phosphorothioates, and at the oxygen of crotonamide carbonyl in crotonamide phosphates. Bulkiness of the substituents at carbon-6 of the pyrimidinyls and/or N-substituents of crotonamides was the steric structural component that contributed to superiority of those OPI for reducing embryonic NAD⁺ levels. The findings of this study provided evidence for the cause-and-effect relationship between yolk sac membrane KFase inhibition and reduced embryo NAD⁺ content in NAD-associated OPI-induced teratogenesis in chickens.

Niraj et al. (2015) developed novel QSAR models using 2D-QSAR and 3D-QSAR with CoMFA methodology for prediction of insecticidal activity of organophosphates (OPs) (Fig. 19). The models were validated with an entirely different external dataset of in-house generated combinatorial library of OPs employing molecular docking against the target AChE protein of *Musca domestica*. The obtained dock scores were in good correlation with 2D-QSAR and 3D-QSAR with CoMFA predicted activities and had the correlation coefficients of 0.62 and 0.63, respectively. The activities predicted by 2D-QSAR and 3D-QSAR with

Fig. 19 Parent structure of OP analogues taken for QSAR analyses Niraj et al. (2015)



CoMFA were also observed to be highly correlated with the value of 0.82. Niraj et al. (2015) screened the combinatorial library molecules for toxicity in non-target organisms and degradability using USEPA-EPI Suite.

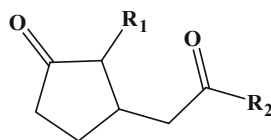
7.5 QSAR of Virucides

A 3D-QSAR model was explored by Zhao et al. (2006) based on the antiviral activity of α -substituted-1, 2, 3-thiadiazoleacetamides tested in vitro against tobacco mosaic virus (TMV). Analysing the QSAR model, the authors proposed that having the same linker between 1,2,3-thiadiazole and benzene ring, compounds that were substituted by solely halogen atom at the 2- or 4-positions of benzene ring had significant potency against TMV, however if a compound was substituted by two halogen atoms at both 2- and 5-positions of benzene ring it had hardly any inhibition. Replacement of the halogen with an electron-donating group, such as methyl or methoxyl, also abolished antiviral activity. Again, exchange of the oxygen linker with sulphur leads to slight loss in activity. After sulfur was oxidized, the bioassay indicated that the antiviral activity of the thioether was slightly better than sulfoxide, whereas worse than sulfone.

7.6 QSAR of Plant Growth Regulators

In the beginning of the 1960s, Hansch et al. (1962) provided momentum to QSAR research by using Hammett constants and hydrophobicity parameters to develop correlation models on plant growth regulators. The significant outcome from the developed equations showed that an increase in activity of the substituted pethoxyacetic acid occurred with an increase in Hammett sigma (σ , an electronic parameter) or as the partition coefficient ($\log P$) was increased, sigma being held constant. Interestingly, the activity values moved to zero with higher values of $\log P$, where the values for sigma were essentially constant. The way in which the two halogens (fluorine and chlorine) affect the electron density at the *ortho* positions by resonance was, however, quite different especially in the example of fluorine. The dramatic difference in activity between the 3- and 4-fluorophenoxyacetic acids seems best explained in terms of the different electron densities at the *ortho* positions.

Fig. 20 Basic scaffold of amino acid conjugate jasmonic acid derivative



Correlation of the biological activity of plant growth regulators and Chloromyces derivatives had been performed with Hammett constants and Partition Coefficients by Hansch et al. (1963). A mathematical equation was developed employing two experimentally based variables, σ and π , for correlating the effect of a given substituent on the biological activity of a parent compound; where, σ is the Hammett substituent constant and π is an analogous constant representing the difference in the logarithms of the partition coefficients of the substituted and unsubstituted compounds ($\pi = \log P_x - \log P_H$). Utilizing π and σ , it became possible to disentangle three of the most important parameters governing the biological activity of organic compounds: steric, electronic, and rate of penetration.

A QSPR analysis was performed with 59 amino acid conjugates of jasmonic acid (Fig. 20) with lipophilicity ($\log P$) parameter by Li et al. (2009a). Statistically significant 2D-QSPR model ($R^2 = 0.990$, $Q^2 = 0.987$) developed by the GFA method showed that the calculated $\log P$ of amino acid conjugates of jasmonic acid was influenced by structural descriptors (Hond Donor and Density), electronic descriptor (Apol) and E-State-keys (S_ssO). Followed by 2D-QSPR model, the results were further compared with the 3D-QSPR model ($R^2 = 0.922$, $Q^2 = 0.841$) with good stability and predictability generated by the MFA-GFA method to study the structural requirements for the $\log P$ of these compounds, indicating that bulky substituents at C7 position were unfavored to the $\log P$ and both steric and electrostatic substituents at C3 position were important to $\log P$. The derived QSPR models were significant to evaluate the important regions and the molecular structural parameters for these compounds.

Li et al. (2009b) applied QSAR analysis to 18 jasmonates and related compounds to study the relationships between chemical structure and their biological activity in barley and tomato bioassay. Statistically significant 2D-QSAR models ($R^2 > 0.880$ and $Q^2 > 0.820$) were developed by GFA, indicating that the biological activity (pKi-1) was principally influenced by thermodynamic, electronic and spatial descriptor and the biological activity (pKi-2) was principally governed by electronic, structural and thermodynamic. Additionally, the authors employed molecular field analysis (MFA) merged with G/PLS method to derive 3D-QSAR models investigating the substitutional requirements for the favorable receptor-drug interaction, and quantitatively indicating the important regions of molecules for their activity. The 3D-QSAR models showed that electrostatic interactions were crucial to pKi-1, and moderate steric interactions were favored to pKi-2, indicating that a tiny transformation in the meso-position and para-position of cyclopentanone may greatly change the biological activity.

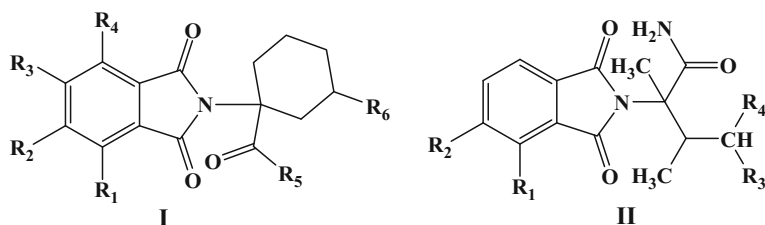


Fig. 21 Two different scaffolds of NSPs employed by Li et al. (2015)

Molecular docking and molecular dynamics simulations were explored to identify the mode of interaction between N-substituted phthalimides (NSPs) and the gibberellin (GA) receptor *GID1A* in order to clarify the relationship between structure and GA-like activity in the NSPs by Li et al. (2015). The obtained results demonstrate that both a multiple-hydrogen-bond network and a ‘hat-shaped’ hydrophobic interaction played imperative roles in the binding of the NSPs to *GID1A* (Fig. 21). The carbonyl group of a phthalimide fragment in the NSPs acted in a comparable manner to the pharmacophore group 6-COOH in GAs, forming multiple hydrogen-bond interactions with residues Gly115, Ser191 and Tyr322 in the binding domain of *GID1A*. The 3D-QSAR study with CoMFA and CoMSIA analysis also confirmed that the GA-like activity of these NSPs was strongly linked to a few H-bond donor and acceptor field contributions of the NSPs to the H-bond interactions with *GID1A*. The outcome showed that increasing the H-bond donor character at the R_5 position was more likely to lead to NSPs with GA-like activity. The introduction of a hydrophobic group at the R_2 position identified to assist the binding of the NSP to the GA receptor. The authors finally designed five new NSP molecules using the binding domain of *GID1A* and then docked into the receptor. Interestingly, two of them showed to have good docking scores due to enhanced hydrophobic contact.

8 Successful Application of QSAR in Food Science

8.1 QSAR of Food Products and Food Supplements

Tilaoui et al. (2007) investigated the performance of the TOPKAT software to predict the Lowest Observed Adverse Effect Level (LOAEL) for more than 600 food-borne chemicals with experimentally established LOAELs, validated by toxicology experts. The reliability of a given prediction was evaluated by the software internal checking OPS algorithm. The models of classification and prediction of the LOAEL were developed with a good estimate of the statistical significance. The set of 2D autocorrelation descriptors correlates well with experimental LOAELs. This approach was usable and satisfactory in the context of chemical food safety

assessments and prioritisation of levels of concern in chemical food safety. It should be kept in mind that the determination of a LOAEL—is still a very challenging endpoint in terms of QSAR-based computer predictions. This is due to a number of limitations, including the lack of scientifically valid and reliable information, as well as the non-specificity of the endpoint. Therefore, the outcome of such predictions should be carefully scrutinised and validated. Interestingly, the prediction of the LOAELs with TOPKAT was reliable for 1/3 of the available compounds (33%) in this study.

Fumonisin B1 (FB1), a *Fusarium* mycotoxin, has received substantial attention from food regulatory agencies due to its prominent immunotoxic, neurotoxic, hepatotoxic, nephrotoxic and carcinogenic properties in animals. In this background, Dambolena et al. (2011) demonstrated a QSAR study concerning the antimycotoxigenic activity of natural phenolic compounds in order to evaluate which molecular properties were important in antifumonisin activity. The obtained results indicated that lipophilicity was the key step for the target molecule to be reached inside the fungal cells. Furthermore, the molar refractivity and saturated area demonstrated the importance of the interactions with specific enzymes, metabolite pools, or signaling pathways. The model obtained from the QSAR analysis can be used to predict the antifumonisin activity of other structurally related molecules and the findings could provide an important contribution in the search for new compounds with antifumonisin activity.

Gu et al. (2011) evaluated systematically the probable of major food proteins as precursors of angiotensin converting enzyme (ACE) inhibitory peptides using QSAR-aided in silico approach followed by demonstrated rationale for choosing the appropriate substrate proteins in preparing ACE inhibitory peptides. Proteins from 15 common food commodities by thermolysin generated 5709 peptides ranging from 2 to 6 amino acid residues were systematically studied as the potential precursors of ACE inhibitory peptides. The results showed that meat proteins from pork, beef and chicken contain the highest number of potent peptides (IC₅₀b10 M), followed by proteins from egg, soybean and canola, whereas proteins from fish (with the exception of salmon) and cereals (oat and barley) contain the least number of peptides. The authors also reported that the release of peptides by in silico digestion might be different from experimental condition where the release of peptides could be affected by a number of factors including the state of the substrate, temperature, pH and specificity of enzyme. The authors strongly concluded that predicted peptides could be released with carefully manipulated digestion conditions.

Dambolena et al. (2012) performed a QSAR study for the inhibition of *Fusarium verticillioides* growth by ten natural phenolic compounds. The results of the experimental determinations demonstrated that in terms of the antifungal activity of natural phenolic compounds on *F. verticillioides*, the following order was found: carvacrol > thymol > isoeugenol > eugenol > vanillin > creosol > m-cresol > o-cresol, p-cresol, and guaiacol. Lipophilicity, Molar refractivity and saturated area were found to be the descriptors that best explained the antifungal activity of these compounds from the mathematical equation. These models could be used in future

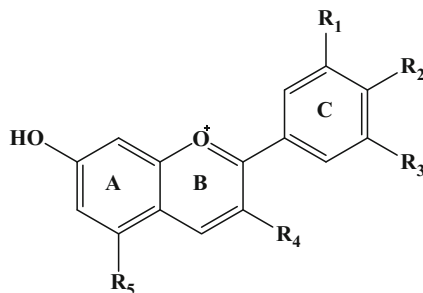
to predict the activity of new compounds and to guide the search for the synthesis of phenolic compounds with the capacity to inhibit *F. verticillioides* growth.

The structural diversity and intrinsic relationship between the multiple bioactivities of 8000 tripeptides have been explored by Wang et al. (2013) by using various molecular modeling techniques, including statistical test, crystal structural examination, binding energetic analysis and QSAR modeling. The authors reported that the first three C-terminal residues were sufficient for a peptide to bind tightly with the active site of ACE protein to exert significant antihypertensive efficacy. A systematical investigation of all possible tripeptides further revealed that there is a good relationship between their ACE-inhibitory potency and antioxidative activity, but unlike dipeptides, both these two desirable properties exhibited insignificant correlations with the undesirable bitterness. This finding suggested that the structural space of tripeptides is sufficiently diverse that makes it possible to achieve a good compromise between multiple bioactivities in a single molecular entity simultaneously.

Predictive QSAR models correlating peptide's structural features with their multi-bioactivities and bitter taste were established at both sequence and structure levels by Tan et al. (2013). Thereafter, the authors used the models to conduct extrapolation on thousands of randomly generated, structurally diverse peptides with chain lengths ranging from two to six amino acid residues. Based on the statistical results obtained from QSAR modelling, the relationship between the antihypertensive activity and bitter taste of peptides at different sequence lengths was investigated in detail by Tan et al. (2013). Moreover, the structural basis, energetic property and biological implication underlying peptide interactions with ACE were analysed at a complex 3D structure level by employing a high-level hybrid quantum mechanics/molecular mechanics scheme. They found the following outcome from the study: (a) bitter taste is highly dependent on peptide length, whereas ACE inhibitory potency has only a modest correlation with the length, (b) dipeptides and tripeptides perform a moderate relationship between their ACE inhibition and bitterness, but the relationship could not be observed for those peptides of more than three amino acid residues and (c) the increase in sequence length does not cause peptides to exhibit substantial enhancement of antihypertensive activity; this is particularly significant for longer peptides such as pentapeptides and hexapeptides.

Zhou et al. (2013) explored the intrinsic relationship between the ACE inhibition and bitterness of short peptides in the scaffold of computational peptidology, attempting to find out the appropriate properties for functional food peptides with satisfactory bioactivities. The QSAR model revealed a significant positive correlation between the ACE inhibition and bitterness of dipeptides, but this correlation was quite modest for tripeptides and, particularly, tetrapeptides. Moreover, quantum/molecular mechanics analysis of the structural basis and energetic profile involved in ACE-peptide complexes unravels that peptides of up to 4 amino acids long are sufficient to have efficient binding to ACE, and more additional residues do not bring with substantial enhancement in their ACE-binding affinity and, thus, antihypertensive capability. The authors came to the conclusion that the tripeptides

Fig. 22 Parent structure of anthocyanin derivative



and tetrapeptides could be considered as ideal candidates for seeking potential functional food additives with both high antihypertensive activity and low bitterness.

Jing et al. (2014) established 3D-QSAR models from 21 anthocyanins (Fig. 22) based on their oxygen radical absorbing capacity (ORAC) and further applied to predict anthocyanins in eggplant and radish for their ORAC values. The contour map results suggested that structural characteristics of anthocyanins favourable for the high ORAC. Four anthocyanins from eggplant and radish were also screened based on the QSAR models. Pelargonidin-3-[(6''-p-coumaroyl)-glucosyl(2 → 1) glucoside]-5-(6''-malonyl)-glucoside, delphinidin-3-rutinoside-5-glucoside, and delphinidin-3-[(4''-p-coumaroyl)-rhamnosyl(1 → 6)glucoside]-5-glucoside potential with high ORAC based on the QSAR models were isolated by Jing et al. (2014) and confirmed their relatively high antioxidant ability. Three key points were concluded by the authors on anthocyanin structure–ORAC relationships. First, a bulky and/or electron-donating substituent at the 3-position in the C ring appears to be necessary for enhancing ORAC of anthocyanins. Additionally, the presence of additional electron-donating and/or hydrophobic groups around the glycosylation might enhance the radical scavenging activity. Lastly, the presence of a hydrogen bond donor group/electron donating group at the R₄ position in the B ring might enhance the radical scavenging activity of anthocyanins.

Vinholes et al. (2014) developed 3D-QSAR models of the hepatoprotective activity of sesquiterpenoids with different backbone structures using an in vitro model system. The developed models allowed the extraction of relevant information suggesting that sesquiterpenoids possessing more compact molecular structures ((-)- α -neoclovene and (-)- α -copaene), low ramification (trans- β -farnesene, trans,trans-farnesol and cis-nerolidol) and less symmetric (according to Gm-total symmetry of the molecule) (trans- β -farnesene, trans,trans-farnesol, (-)- α -copaene, cis-nerolidol and (-)- α -neoclovene) will be more effective for endogenous hepatoprotection. Additionally, compounds with electronegative substituents (guaiazulene, trans,trans-farnesol, trans- β -farnesene, (+)-valencene and (-)- α -copaene), less ramified structures (trans,trans-farnesol, trans- β -farnesene, (-)- α -copaene and guaiazulene) and with more symmetry (according to G2e-symmetry considering the second component) with an electronegative terminal fragment (guaiazulene, trans- β -farnesene and trans,trans-farnesol) seem to be more effective for the induced hepatoprotection as suggested by Vinholes et al. (2014). This study supports the

existing tendencies of valorisation of natural products as a source of bioactive compounds for the formulation of foods and/or nutraceuticals enriched extracts.

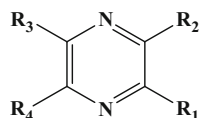
Takaki et al. (2015) estimated metabolic rate using existing QSAR biodegradation models of microorganisms (BioWIN) and fish (EPI-HL and IFS-HL). Then they incorporated the obtained simulated metabolic rate into the mechanistic cattle biotransfer models (RAIDAR, ACC-HUMAN, OMEGA, and CKow). The goodness of fit tests showed that RAIDAR, ACC-HUMAN, OMEGA model performances were significantly improved using either of the QSARs when comparing the new model outputs to observed data. The CKow model was the only one that separates the processes in the gut and liver. The developed model showed the lowest residual error of all the models tested when the BioWIN model was used to represent the ruminant metabolic process in the gut and the two fish QSARs were used to represent the metabolic process in the liver. The testing included EUSES and CalTOX which are KOW-regression models that are widely used in regulatory assessment. New regression models based on the simulated rate of the two metabolic processes were also proposed as an alternative to KOW-regression models for screening risk assessment. The authors stated that the modified CKow model is more physiologically realistic, but has equivalent usability to existing KOW-regression models for estimating cattle biotransfer of organic pollutants.

Kiwamoto et al. (2015) aimed to develop physiologically based kinetic/dynamic (PBK/D) models to examine dose-dependent detoxification and DNA adduct formation of a group of 18 food-borne acyclic α,β -unsaturated aldehydes without 2- or 3-alkylation and with no more than one conjugated double bond. The PBK/D models were obtained using a training set of six aldehydes. Using the developed QSAR equation, PBK/D models for the other 12 aldehydes were defined. The results revealed that DNA adduct formation in the liver increases with decreasing bulkiness of the molecule especially due to less efficient detoxification. 2-Propenal (acrolein) was identified to induce the highest DNA adduct levels. The authors concluded that at realistic dietary intake, the predicted DNA adduct levels for all aldehydes were two orders of magnitude lower than endogenous background levels observed in disease free human liver, suggesting that for all 18 aldehydes DNA adduct formation is negligible at the relevant levels of dietary intake. The present study elucidated a possible negligible DNA adduct formation in the liver upon oral exposure to a range of acyclic food-borne α,β -unsaturated aldehydes at relevant levels of dietary intake and provided a proof of principle for the use of QSAR based PBK/D modelling to facilitate group evaluations and read-across in risk assessment.

8.2 QSAR of Flavor of Food Products and Food Supplements

Buchbauer et al. (2000) constructed qualitative models to explain the aroma of 46 bell-pepper flavor from pyrazines scaffold (Fig. 23) using the graphic features of

Fig. 23 General structure of Pyrazines derivative used by Buchbauer et al. (2000)

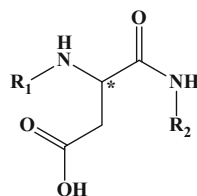


CoMFA in combination with the “classical” QSAR analysis performed by MLR. Biological activities were used in terms of the detection threshold value of the aroma compound in water. The CoMFA analysis concluded that the bulky groups in the substituent positions R₂ and R₁ increase the biological activity, whereas the unfavorable contribution of bulky groups in the substituent positions R₃ and R₄. Moreover, the equation suggested that increased positive charges at the first atoms of substituent R₃ ($C(R_3)$) were of advantage for bell-pepper flavor, because their values were mainly positive and had positive regression coefficients. The unfavorable effect of a negative electrostatic field in the region of substituent R₂ resulted from the values of CR_2 which were negative, and the regression coefficient which was positive, indicating that the more negative CR_2 was, the lower the contribution to “bell-pepper flavor” will be. This was also in agreement with the CoMFA picture, where a positive field at R₂ favors the biological activity. The molecular surface of the molecules should not be too high, because it caused to a decrease of the biological activity. The last significant descriptor turned out to be the log P value of substituent R₁. The favorable effect of low log P values at R₁ was suggested by the MLR regression analysis.

Tromelin and Guichard (2003) investigated a 3D-QSAR study using Catalyst software to explain the nature of interactions between flavor compounds and β -lactoglobulin. For this purpose, a set of 35 compounds were chosen with dissociation constant values previously determined by affinity chromatography. An automated hypothesis generation using the HypoGen software produced a model that made a precious inference of affinity and provided a clarification for the lack of correlation previously observed between the hydrophobicity of terpenes and the affinity for the protein. The outcome of the model suggested that aroma binding to β -lactoglobulin was caused by both hydrophobic interactions and hydrogen bonding, which played a critical role. This observation provided an explanation for the observed binding constants, which were not in relation to the molecules’ hydrophobicities. It was important to note that the hydrophobicity was not the only important feature; the topology of the hydrocarbon chain and hydrogen bonding were also essential for the predictive capability of the model.

A QSPR approach was explored to evaluate the influence of the chemical structure of aqueous matrixes over the partition coefficient between the gas phase and the matrix by Chana et al. (2006). The determination of the partition coefficient of flavor ingredients was performed by headspace analysis at equilibrium for both saline solution and *t*-carrageenan gel. The QSAR equation was generated by the GFA method available in the Cerius2 package. The best obtained equation involved only five descriptors, which encode electronic properties of charges repartition on the molecule (Jurs-RNCS and Dipole-Z) and molecules’ shapes (PMI-Y,

Fig. 24 Structure of 3-Amino-succinamic acid derivatives used by Tarko et al. (2006)



Shadow-XY, and RadOf-Gyration), both for saline solution and for *ι*-carrageenan gel. However, the best-fitting equation for carrageenan gel was obtained with a quadratic relation, suggesting that the effect of carrageenan polymers only modulates, but it did not change the interaction of aroma compounds with water molecules. In case of molecular orientation and alignment, the greater the chain length and moderate the branching, the larger are the values of the three shape descriptors. Such results showed that the higher the charge more is the retention, and the more globular form or the more ramifications within the molecule, less is the retention for esters and ketones. Such behavior was consistent with the nature of water with a high ionic force, where hydrophobic compounds, usually highly branched and/or with long aliphatic chains, were barely soluble.

Tarko et al. (2006) performed QSAR studies with the PRECLAV computer program using a database containing 136 3-amino-succinamic acid derivatives (Fig. 24). The developed model suggested that the virtual molecular fragments that lead to a significant increase of the sweetness power (SP) were $-\text{CN}$ (cyano) and $-\text{C}_6\text{H}_4-\text{NHCONH}-$ (aryl-substituted urea). The non-conjugated or weakly conjugated virtual fragment $-\text{NH}_2$ leads to a significant decrease of the SP value. Additionally, the SP was favorably influenced by the size of the molecule. The linear functions of the descriptors strongly described the SP of the studied derivatives.

Wu and Aluko (2007) applied QSAR modeling as a tool to determine the type and position of amino acids that contribute to bitterness of di- and tri-peptides. Datasets of bitter di- and tri-peptides were constructed followed by modeling using PLS regression based on the three z -scores of 20 coded amino acids. The results showed that a single-component model could explain 52% and 50% of the bitterness threshold of bitter di- and tri-peptides, respectively. The PLS model determined that hydrophobic amino acids at the carboxyl-terminus and bulky amino acid residues adjacent to the carboxyl terminal were the major determinants of the intensity of bitterness of di- and tri-peptides. However, there was no significant correlation observed between bitterness of di- and tri-peptides and their angiotensin I-converting enzyme-inhibitory properties. Knowledge of the type and position of amino acids that contribute to bitterness could provide a basis for elimination of certain residues from food proteins or rearrangement of residues on the primary structure using genetic engineering techniques. The results can also enhance the production of less bitter protein hydrolysates through appropriate choice of

enzymes that cleave the bonds between the amino acid residues shown to be important determinants of peptide bitterness.

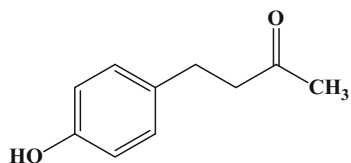
A series of aspartame analogues and chemically modified derivatives was evaluated by Vepuri et al. (2007) as artificial sweeteners using 3D-QSAR with GFA, CoMFA and CoMSIA analysis. The combination of steric, electrostatic, hydrophobic, and H-bond acceptor fields in CoMSIA gave better results than CoMFA model. The predictive ability of 3D-QSAR with GFA, CoMFA, and CoMSIA were determined using a test set giving predictive correlation coefficients of 0.375, 0.535, and 0.596, respectively, indicating better predictivity of CoMSIA compared to the other methods. Combining the outcome of all analysis, the authors stated following findings from the study:

- Substitution of bulkier groups at the amino terminal end of aspartic acid could increase the sweetness value.
- Conversely the presence of a bulky group at the oxygen atom of ester linkage could decrease sweetness.
- The presence of positively charged groups at the amide linkage and negatively charged groups near the ester linkage should increase the sweetness potency value.
- Presence of bulky and hydrophobic groups at the amino terminal end of aspartic acid should increase the sweetness to a greater value.
- Hydrogen-bond acceptor groups at the carbon atom attached to nitrogen of the amide group should further increase the sweetness. The higher sweetness value for the ester derivatives of aspartic acid compared to non-ester derivatives was due to the H-bond acceptor nature of C=O in the ester group.

Yang et al. (2011) constructed three QSAR models for the prediction of sweetness of 103 compounds. The molecules were represented by three descriptors. On the basis of the Kohonen's Self-Organising Neural Network (KohNN) map, the whole data set was split into a training set including 58 compounds and a test set including 45 compounds. Then, logarithmic value of sweetness (logSw) was predicted by using MLRs, ANN and SVM regression analyses. For the test set, the correlation coefficient of 0.925, 0.932 and 0.943 for the MLR, ANN and SVM, respectively, were achieved. This work revealed that XlogP, 2DACorr_Ident(2)/2DACorr_Polariz(0) and SurfACorr_HPP_6 determined the sweetness of the studied compounds. The authors also concluded that nonlinear methods such as SVM and ANN provided better models than MLR analysis; therefore nonlinear methods were preferable in the modelling of sweetness.

Two quantitative models were built by Zhong et al. (2013) to predict the logSw (the logarithm of sweetness) of 320 unique compounds with a molecular weight from 132 to 1287 and sweetness from 22 to 22,500,000. The QSAR models were built employing MLR and SVM analysis. For the test set, the correlation coefficients of 0.87 and 0.88 were obtained by MLR and SVM, respectively. The model consisted of twelve descriptors (including 2DACorr_Sigchg_2/2DACorr_Polariz_0, LogS, 3DACorr_PiChg_5, and 9 RDF descriptors). The selected molecular

Fig. 25 Structure of Raspberry ketone



descriptors predicted the sweetness statistically and they could interpret the sweet taste system theory of a sweetener, such as the AH/B System Theory founded by Shallenberger and Acree.

Raspberry ketone (4-(4-hydroxyphenyl)-2-butanone) (Fig. 25) is known as a food supplement. Due to the limited toxicological information on raspberry ketone, Bredsdorff et al. (2015) performed an *in silico* profiling of raspberry ketone by QSAR models. The QSAR screening was performed in the Leadscape Model Applier Version 1.7.4 software (<http://www.leadscape.com/>). A total of 54 commercial models from 4 suites developed in a collaboration with the U.S. Food and Drug Administration were run: Developmental Toxicity Suite (27 models for Rats, Mice, Rabbits and overall Rodents related to foetal growth retardation, foetal weight decrease, foetal death, post-implantation loss and pre-implantation loss), Reproductive Toxicity Suite (9 models for Rats, Mice and overall Rodents related to overall reproductive effects and sperm effects), Human Adverse Hepatobiliary Effects Suite (5 models), and Human Adverse Cardiological Effects Suite (13 models). In conclusion, the results from QSAR models were mostly negative and only points towards potential hazards, however, considering the available data it cannot be excluded that raspberry ketone has potential adverse effects on reproduction or development or was cardiotoxic.

A predictive QSPR was developed by Rojas et al. (2015) for modeling the retention index measured on the OV-101 glass capillary gas chromatography column, in a set of 1208 flavor and fragrance compounds. Compounds were simultaneously analyzed through MLR analysis using the replacement method (RM) variable subset selection technique. The authors performed the modeling in three steps, the first one by considering all descriptor blocks, the second one by excluding conformational descriptors blocks, and the last one by analyzing only 3D-descriptors families. The results clearly suggested that 3D-descriptors do not offer relevant information for modeling the retention index, while a topological index such as the solvation connectivity index of first order had a high relevance for prediction. The authors wrapped up the discussion suggesting that the conformation-independent QSPR method can emerge as an alternative approach for developing models based on constitutional and topological molecular features of compounds.

Rojas et al. (2016) developed predictive QSPR model for natural and synthetic sweeteners in order to predict and model relative sweetness (RS). The data set was composed of 233 sweeteners collected from diverse sources of literature and a total of 3763 non-conformational Dragon molecular descriptors were calculated which

were simultaneously analyzed through MLR analysis coupled with the replacement method variable subset selection technique. The established six-parameter model was validated through the cross-validation techniques, together with Y-randomization and applicability domain analysis. The important finding of the study was that the presence of hydrophobicity in a sweetener permits a favorable partition of the substance between the aqueous saliva fluid and the lipidic taste receptor membrane. The hydrophilicity of a sweetener allows its diffusion through the saliva to rapidly interact with the taste receptor.

8.3 QSAR of Antioxidants

Various classes of chemical entities have been found to exhibit antioxidant activity. A variation in their substitution pattern of these antioxidant molecules imparts a difference in their physicochemical properties, which in turn influences their reactivity with toxic free radicals. During the last decade, diverse classes of such chemical entities have gained remarkable attention from different groups of medicinal chemists for their ability to scavenge free radicals, and thus inhibit systemic damages like lipid peroxidation. Successful QSAR models developed by different groups of authors based on the available diverse classes of antioxidant molecules are discussed below.

8.3.1 QSAR of Phenolic Antioxidants

Antioxidant activity of wine polyphenols was modeled by Rastija and Medic-Saric (2009) using the QSAR technique with the descriptors calculated from 2D and 3D representation of the molecules. The significant models for the antioxidant activity of the polyphenols showed that the zero-order connectivity index (${}^0\chi$) and molar refractivity were the useful parameters for modeling free radical scavenging activity of polyphenols belonging to different groups (phenolic acids and flavonoids, flavans, flavonols and stilbene). The models thus developed also indicated that lipophilicity and van der Waals volume were the significant molecular descriptors for prediction of antioxidant activity of flavonoids in the lipophilic phase. They also demonstrated that the number and the arrangement of free hydroxyl groups on the flavonoid skeleton, or on the phenol ring of phenolic acids together with the shape, size, mass and steric properties of the molecules bear considerable effects on the activity profile of these molecules.

Cheng et al. (2002) developed QSAR models for studying multiple mechanisms underlying the reaction between hydroxyl radical and phenolic compounds and reported that the reaction rate constant (K_S) bears good correlation with hydroxyl O–H bond strength, electron-donating ability [ionization potential approximated by HOMO energy level], enthalpy of single electron transfer, and spin distribution of phenoxyl radicals after H-abstraction. MLR analysis indicated that, in addition to

H-atom transfer, electron transfer process and stability of the resulting phenoxyl radicals also significantly influence the reactivity of quenching hydroxyl radicals.

Modeling and statistical analysis for DPPH free radical scavenging activity of phenolic compounds was performed by Velkov et al. (2007). The authors reported that there existed a significant linear correlation between the free radical scavenging activity and the spin density as well as the HOMO energy of the molecules, and inferred that the radical scavenging activity of the phenolic compounds is efficiently influenced by the electron donor ability of the O–H group to the aromatic ring, the occurrence of substituents with positive mesomeric and inductive electronic effects and the presence of hydrogen bonds involving dissociable hydroxyl group (DHG) and adjacent functional groups.

Ray et al. (2008) performed QSAR studies in order to predict the lipid peroxidation inhibition potential of some phenolic antioxidants in phosphate buffered and pre-emulsified linoleic acid systems. The models were built in this study using the stepwise regression and MLR/with factor analysis (FA) as the data processing step for variable selection (FA-MLR), and it was revealed that the bond dissociation enthalpy of the O-H bond and the MAXDP (maximal electrotopological positive variation) descriptor bear negative influences on the lipid peroxidation inhibition potency of these molecules.

Mitra et al. (2011) explored QSAR models for 33 phenolic derivatives bearing NO donor groups. These models chiefly inferred that presence of substituted aromatic carbons, long chain branched substituents, an oxadiazole-N-oxide ring with an electronegative atom containing group substituted at the 5 position of the parent nucleus, increase in the positively charged surface area and the volume of the molecules favours the antioxidant activity profile of these compounds. Long chain branched substituents lacking symmetry about the centre of mass of the molecule exhibit improved antioxidant activity. The authors extended the work through design of 15 new molecules of this class with subsequent *in silico* prediction of their activity based on the developed models.

Chen et al. (2015) investigated the structure-thermodynamics-antioxidant relationships of 20 natural phenolic acids and derivatives using DPPH[•] scavenging assay, DFT calculations at the B3LYP/6-311++G(d,p) levels of theory, and QSAR modeling. The authors explored three main working mechanisms: hydrogen atom transfer (HAT), electron transfer-proton transfer (SETPT) and sequential proton loss-electron transfer (SPLET) in four micro-environments (gas-phase, benzene, water and ethanol). Subsequently, computed thermodynamics parameters (BDE, IP, PDE, PA and ETE) were compared with the experimental radical scavenging activities against DPPH[•]. Combined theoretical and experimental investigations demonstrated that the extended delocalization and intra-molecular hydrogen bonds were the two main contributions to the stability of the radicals. The C=O or C=C in COOH, COOR, C=CCOOH and C=CCOOR groups, and orthodiphenolic functionalities were shown to favorably stabilize the specific radical species to enhance the radical scavenging activities, while the presence of the single OH in the *ortho* position of the COOH group disfavors the activities. The authors concluded from

the study that HAT was the thermodynamically preferred mechanism in the gas phase and benzene, whereas SPLET in water and ethanol.

8.3.2 QSAR of Flavonoids Exhibiting Antioxidant Property

Farkas et al. (2004) studied 36 flavonoids using PLS projection of latent structures method and developed significant QSAR model with several constitutional descriptors, 2D topological and connectivity indices. They reported plots for PLS component scores indicate that the model provides a suitable prediction for most of the flavonoids, and since the model was developed for antioxidant activities of a diverse set of flavonoids, the model could be used for classification of different flavonoid groups.

A series of flavonoids were investigated by Rackova et al. (2005) in order to examine the structural parameters contributing to the antilipid peroxidative activity. The significant QSAR models developed by them indicated the importance of the electronic parameters, viz. hydration energy and energy of LUMO for the lipid peroxidation inhibitory potential of these flavonoids illustrate the hydrophilic and electrophilic properties of the molecules respectively, and inferred that the highest (absolute) values of E_{HYDR} were obtained for most of the potent flavonoids (possessing the highest number of OH groups), while the lowest (absolute) values of E_{HYDR} were attributed to flavonoids that exerted low antioxidant activity.

Durand et al. (2007) performed a 2D QSAR analysis of flavonoids and hexahydropyridoinoles in order to predict the antioxidant activity of pinoline derivatives (1,2,3,4-tetrahydro- β -carboline). The work highlights that the antioxidant activity of various classes of compounds is also governed by topological and functional group parameters. In another study, Calgarotto et al. (2007) performed multivariate study on flavonoids with the aim to select electronic properties responsible for their peroxynitrite scavenging activity. The authors reported that higher HOMO energy for the flavonoids compounds facilitates transfer of the reducing electrons to the peroxynitrite radical, while atomic charges associated with the hydroxyl groups are involved in the electron transfer process between the flavonoids and the peroxynitrite radical.

Mitra et al. (2010a) studied a series of isoflavones, isoflavanes and biphenyl ketones derivatives (Fig. 26) which were modeled previously using the Fujita-Ban analysis as well as based on different other chemometric tools to explore the influence of different substituents on the free radical scavenging activity of the molecules. Mitra et al. (2010a) suggested that the presence of hydroxyl substituent at different positions of the A and B rings substantially influences the antioxidant activity of these molecules while a methoxy substituent at R_6 position of the B ring favours activity. An increase in the number of hydroxy substituents enhances the availability of hydrogen bond donor groups on the antioxidant molecules and consequently facilitates the neutralization of toxic free radicals. Besides these, molecules lacking the *ene* fragment of the pyran ring and molecules with an open pyran ring show enhanced activity profile.

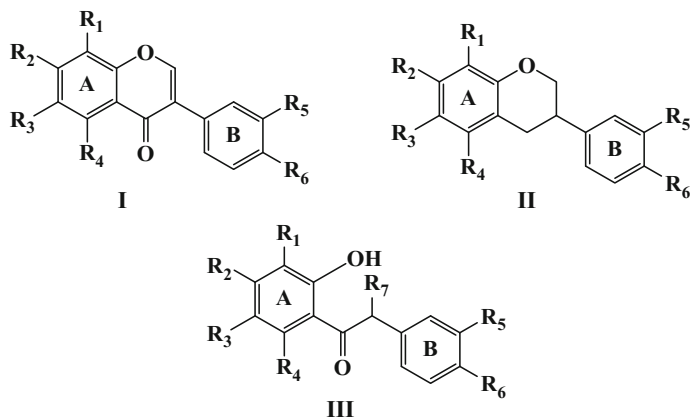
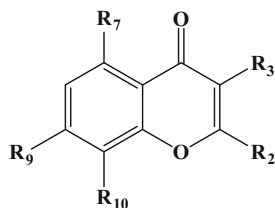


Fig. 26 Common scaffold of three different flavones derivatives

Fig. 27 Parent structure of synthetic chromone derivative as employed by Mitra et al. (2012a)

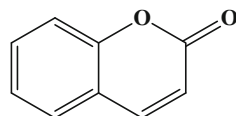


8.3.3 Chromone Derivatives

Samee et al. (2008) developed statistically significant QSAR models for a series of 7-hydroxy, 8-hydroxy and 7, 8-dihydroxy synthetic chromone derivatives for their DPPH free radical scavenging activities using genetic PLS approach for model development. The MFA equation suggested that electronegative group on benzoyl ring and electropositive group on phenyl ring are the important factors controlling the antioxidant activity of these chromone derivatives.

Thirty six synthetic chromone derivatives (Fig. 27) were employed for development of four QSAR models, namely 3D-pharmacophore mapping, CoMSIA, HQSAR and group based QSAR (G-QSAR) techniques by Mitra et al. (2012a). The importance of the hydrogen-bond acceptor feature was revealed by all four analyses. Thus, the hydroxyl substituent at the R_{10} position and the benzoyl substituent at the R_3 position of the chromone nucleus were indispensable fragments for an enhanced antioxidant activity. Additionally, the ketonic group at C4 further enhances the abilities of the molecules to interact with the toxic free radicals through a mechanism of electron transfer followed by deprotonation. The CoMSIA analysis indicated that bulky substituents were disfavored at R_2 position. For the 3D pharmacophore model, the presence of the ring aromatic and the hydrophobic features over the substituents at the R_2 and R_3 positions, respectively, indicated that


Fig. 28 Scaffold of Coumarin derivative



such groups separated by the specific distance of 5.890 Å were essential for the enhanced activities of the molecules. Similar results were also obtained from the CoMSIA study, where these substituents map to the hydrophobically favored contours. Moreover, the HQSAR contour study also revealed the importance of such fragments, with the green color for the substituent at R₃ indicating its maximum contribution. Finally, the G-QSAR models closely matched with those of the remaining models, indicating the important impact of the hydroxyl substitution at the R₁₀ position on the antioxidant activity profiles of the chromone derivatives, in addition to the remaining essential features, such as the presence of the substituted benzoyl fragment at the R₃ position and the substituted aromatic fragment at the R₂ position.

8.3.4 Coumarins

Diverse QSAR approaches had been utilized for identifying the essential structural attributes imparting potential antioxidant activity profile of a series of coumarin derivatives (Fig. 28) by Mitra et al. (2013a). The descriptor based QSAR model provided a quantitative outline regarding the structural prerequisites of the molecules, while 3D pharmacophore and HQSAR models emphasized the favourable spatial arrangement of the various chemical features and the crucial molecular fragments respectively. All the models inferred that the fused benzene ring and the oxygen atom of the pyran ring constituting the parent coumarin nucleus captured the prime pharmacophoric features imparting superior antioxidant activity to the molecules.

Descriptor based QSAR, 3D pharmacophore, HQSAR and G-QSAR have been employed for 50 coumarin derivatives by Mitra et al. (2013b). The descriptor-based QSAR model primarily implicated the importance of ketonic oxygen fragment followed by the secondary amine group and the unsubstituted aromatic carbon fragments. Similar observation signifying the importance of the =O fragment was also obtained from the 3D pharmacophore model as well as the HQSAR and the G-QSAR analyses. Further, the phenyl group substituted to the side chain attached at the C3 position of the parent nucleus implicated to be an important feature based on its mapping with the ring aromatic feature of the 3D pharmacophore model. The  linker obtained from the HQSAR analysis and the =CH- fragment constituting the substituent attached to the C3 position of the coumarin moiety as marked to be essential by the G-QSAR model enabled the molecules to map with the essential ring aromatic feature. The G-QSAR model also implicated the importance of the hydrophobic nature of the substituent at the C4 position of the

parent moiety. The importance of the benzene fragment of the benzopyran ring was also implicated by both the 3D pharmacophore model and the HQSAR analysis. In case of the 3D pharmacophore model, this fragment mapped with the essential hydrophobic feature, while the HQSAR contribution map implicated its importance by marking the bonds and atoms of this fragment.

Ferric-reducing antioxidant power (FRAP) assay values were used to explore QSAR models of 37 antioxidant coumarin derivatives by Erzincan et al. (2015). The authors developed QSAR models employing fully optimized structures by semi-empirical PM6 method using SPARTAN 10 software and descriptors were calculated by DRAGON 6.0 software. The MLR models were developed with QSARINS 2.2.1 software. Robustness, reliability and predictive power of the models were tested by internal and external validations ($R^2 = 0.924$; $RMSE_{TR} = 0.213$; $R_{ext}^2 = 0.887$; $RMSE_{ext} = 0.255$; $CCC_{ext} = 0.939$). Descriptors appeared in the model revealed that complexity, H-bond donor and lipophilic character were important parameters in describing the antioxidant activity. Additionally, the authors also designed 31 new antioxidant coumarin derivatives and predicted their antioxidant activity employing the best two-descriptor model. Interestingly, twelve compounds showed promising predicted antioxidant activity than the studied ones.

8.3.5 QSAR of Miscellaneous Class of Chemicals with Antioxidant Activity

Ancerewicz et al. (1998) explored trimetazidine derivatives, mostly having a free phenolic group, for their radical scavenging and antioxidant properties, and assessed their reaction with DPPH as a measure of radical scavenging capacity. From the various significant QSAR models developed, it was revealed that lipophilicity plays a prime role in the process of inhibition of lipid peroxidation, and hydrogen abstraction was not the sole mechanism responsible for the reaction between antioxidants and radicals produced in the Fenton reaction.

The 3D pharmacophore technique and CoMFA methods were employed by Vajragupta et al. (2000) in order to study the activity of 13 radical scavengers. The classical QSAR models indicated that the electronic parameters together with steric molar refractivity and lipophilicity were the determinant factors contributing to the antioxidant activity of the molecules. Again, the 3D QSAR studies revealed that the structural properties contributing to the activity were not only lipophilic, but also the optimum steric property and geometry of side-chain composition.

Beltran et al. (2007) performed QSAR analysis of a series of di-phenyl-tin^{IV}-salicyliden-ortho-aminophenols with their antioxidant activity values (IC_{50}) calculated based on their ability to inhibit thiobarbituric acid reactive substances (TBARS). They reported that there exists a significant correlation between the TBARS activity and the *ortho* aminophenol substitutions. Besides the Hammett constant (σ), one bond tin coupling constants and tin chemical shift also bear a linear relationship with the activity of these molecules. The authors concluded that

implied molecular variables can become trackers for the calculation of TBARS inhibitory activity.

The QSAR models were developed on the basis of DPPH and ABTS tests and the attack on DNA by hydroxyl radicals, and ANN tool was used for model development by Prouillac et al. (2009). The network comprised of 12 input nodes in the input layer, 1 node in the output layer, with no nodes in the hidden layer and thirty thousand learning cycles. The results from DFT and QSAR studies concluded that thermodynamically thiol derivatives may react more efficiently with hydroxyl radicals than with aminothiols and the importance of the HOMO energy in the SAR strongly suggested the involvement of a hydrogen donation in the free radical scavenging process.

Predictive pharmacophore models developed by Mitra et al. (2010b) for 26 arylamino-substituted benzo [b] thiophenes exhibiting free radical scavenging activity through DPPH radical inhibition assay. The most predictive pharmacophore model developed using the conformers obtained from the *BEST* method consisted of three features: hydrogen bond donor, hydrogen bond acceptor and aromatic ring. The authors extended the study: further pharmacophore hypotheses developed using conformers obtained from the *FAST* method yielded models with good predictivity, with the best one consisting of two features: hydrogen bond donor and hydrogen bond acceptor. Both of the pharmacophores highlighted the importance of HBA and HBD features to the potent antioxidant activities of these molecules. The presence of the secondary amino hydrogen donor group and the electronegative oxygen atom of the methoxy substituent were the prime structural attributes associated with an increased activity profile in this series of substituted benzothiophenes. Besides these features, the pharmacophore developed with conformers generated from the *BEST* method also indicated the influential role of the benzothiophene moiety in modulating the antioxidant activities of these molecules, as implied by the presence of a ring aromatic (RA) feature in the selected pharmacophore. The models may be utilized to estimate the potential antioxidant activities of virtual libraries of newly designed antioxidant molecules of this class prior to synthesis or biological testing.

Li et al. (2011) explored 214 tripeptides containing either His or Tyr residue to study QSAR of antioxidative tripeptides. The PLS tool was employed to model antioxidative tripeptides activities with independent parameters computed from each amino acid, including Divided Physico-chemical Property Scores (DPPS), Hydrophobic, Electronic, Steric, and Hydrogen (HESH), Vectors of Hydrophobic, Steric, and Electronic properties (VHSE), Molecular Surface-Weighted Holistic Invariant Molecular (MS-WHIM), isotropic surface area-electronic charge index (ISA-ECI) and Z-scale. The model concluded that the DPPS was better to describe the amino acid of antioxidative tripeptides revealing that the importance of the center amino acid and the N-terminal amino acid were far more than the importance of the C-terminal amino acid for antioxidative tripeptides. The hydrophobic (positively to activity) and electronic (negatively to activity) properties of the N-terminal amino acid were suggested to play the most important significance towards activity with positive and negative contributions, respectively, followed by

positive contribution of hydrogen bond of the center amino acid. The N-terminal amino acid should be a highly hydrophobic and low electronic amino acid (such as Ala, Gly, Val, and Leu). On the other hand, the center amino acid would be an amino acid that possesses high hydrogen bond property (basic amino acids like Arg, Lys, and His). The obtained structural characteristics of antioxidative peptide were significant to the further research of antioxidative mechanism.

Pérez-Garrido et al. (2012) developed a QSAR model for a heterogeneous group of substances with TOPS-MODE descriptors for an interpretation of their antioxidant activity in the form of bond contributions, which in turn revealed that the prime driving forces for their radical scavenging activity were hydrogen bond donation ability and polarity.

A group of cinnamic acid and caffeic acid derivatives having the ability to inhibit lipid peroxidation composing were modeled by Mitra et al. based on three different techniques: descriptor based QSAR models, 3D pharmacophore models and HQSAR models Mitra et al. (2012b). The results obtained from all the models well corroborate with each other and signify the importance of the ketonic oxygen of the amide/acid fragment and the ethereal oxygen substituted to the parent phenyl ring of the molecules under study.

The importance of the different substituents of 59 azole derivatives (Fig. 29) was quantitatively analyzed employing the descriptor-based QSAR and G-QSAR models, while the pharmacophore, CoMSIA and HQSAR models were employed to identify the prime molecular features accounting for the potent antioxidant activity of the molecules by Mitra et al. (2013c). The descriptor based QSAR model inferred that an increase in hydrophobicity of the azole molecules due to the presence of secondary amine fragments with aliphatic and aromatic substituents together with extensive branching in the molecular structure reduced the activity profile of the molecules. Additionally, an increase in the positively charged surface area of the molecules also disfavoured their activity profile. On the contrary, fragments bearing an aromatic carbon attached to three heteroatoms and a secondary amine linked by single bonds to any two groups were essential for optimum antioxidant activity of the molecules. The 3D pharmacophore mapping and the CoMSIA studies identified the prime molecular features comprising the response pharmacophore of the azole derivatives. The best pharmacophore model implicated the importance of the azole ring together with the aryl substituent at C5 position and the 3-aryl substituted sydonyl fragment attached to the azole moiety at C2 position by a long aliphatic chain. These fragments rightly captured the three ring aromatic

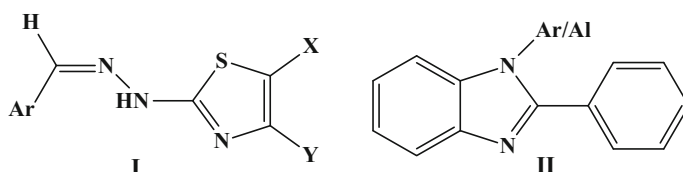


Fig. 29 Structures of azole derivatives

and the hydrophobic features essential for the antioxidant activity of the molecules. This observation rightly matched with the CoMSIA map which bears an electrostatically favoured contour over the azole fragment and the azo linkage indicating their importance for influencing the activity data of the molecules. Additionally, the ability of the phenyl fragment, substituted at the C5 position of the azole ring, to map with the sterically favoured feature in case of the CoMSIA was well corroborated with the pharmacophore analysis which ensured mapping of this fragment with the essential hydrophobic feature. Besides these, the HQSAR model also implicated the importance of the azole moiety as well as the secondary amine fragment linked to the C2 position of the azole ring and the connection of the azole ring to the phenyl fragment attached at its C5 position. Similar results were also obtained from the G-QSAR model which inferred the importance of the five membered azole ring for the optimum activity profile of the molecules. On the contrary, the G-QSAR model also implicated the detrimental effects of the five membered triazole ring, the single bonded thio (-SH) group and the imine (=NH) group present in the side chain of the parent azole ring bearing fragment.

Li and Li (2013) investigated QSAR modeling of antioxidative peptides for scavenging radicals in three free radical systems [Trolox-equivalent antioxidant capacity (TEAC), oxygen radical absorption capacity (ORAC), and superoxide radical (SOR)] and observed a relationship between the physicochemical properties of the C-terminal and N-terminal region and antioxidant potency. It is identified that the properties of amino acids at the C-terminal regions were more important than those in the N-terminal regions for predicting antioxidant activity. Antioxidant activity was correlated with structures related to the electronic, hydrophobic, steric, and hydrogen bonding properties of amino acids at the N-terminal and C-terminal regions, and each terminus covers the three (or four) amino acid residues of peptides containing up to 20 amino acid residues. The properties of amino acids at $C2 > C1$ for TEAC, $C3 > C4 > C1$ for ORAC, and $C4 > C1 > N1$ for SOR were highly correlated with antioxidant activity. Although electronic property most significantly contributed to antioxidant activity in the three free radical systems, it had complex effects at each position. Bulky hydrophobic amino acids at the C-terminal were related to the antioxidant activity of peptides in the three free radical systems. For peptides in the TEAC database, the relationship between the N-terminal segment (N2, N3) and the activity increased when longer peptides were included, which reflects the likely influence of stericity.

Chen et al. (2014) performed a 3D-QSAR study employing CoMSIA on a set of 27 curcumin-like diarylpentanoid analogues (Fig. 30) and their DPPH scavenging activities. The results indicated that a combination of steric, hydrophobic, hydrogen bond donor and hydrogen bond acceptor fields showed good correlative and predictive properties. Moreover, the authors have been able to purposely derive chemical properties which were important to the activity, and hence adopt a rational approach towards the selection of substituent's at various positions in the curcumin scaffold. Additionally, favored and disfavoured regions for enhanced antioxidative activity were suggested. A significant cross-validated correlation coefficient $Q^2 = 0.784$, SEP = 0.042 for CoMSIA was obtained, indicating the statistical

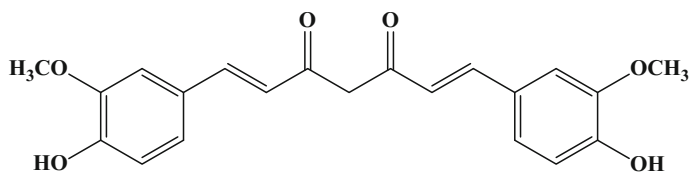


Fig. 30 Structure of the curcumin

significance of the correlation scaffold. The result can be used as a guide to design compounds that potentially have better activity against oxidative damage.

The extensive literature search related to the development of in silico models, especially QSAR, for agrochemicals, food product, peptides, antioxidants, flavors revealed that a wide variety of chemicals have been analysed for determining the structural attributes of the molecules regulating their response profile. However, there still lies a range of chemical entities that did not attain the significant amount of analysis in this framework. Analyzing the immense importance of molecular modeling, the present book chapter supports more modeling work for molecules belonging to a wide variety of chemical classes from the vast area of practical fields of agricultural and food interest.

9 Databases

In the food chemistry field, a number of databases have been compiled; though in certain cases, food components in databases are not single chemicals, but rather mixtures (Martinez-Mayorga and Medina-Franco 2009). A number of food, phytochemicals and flavor-related molecular databases are available in the present time. We have made a comprehensive presentation for chemoinformatic characterization of a subset of the Flavor and Extract Manufacturers Association (FEMA); Generally Recognized As Safe (GRAS) list of approved flavoring substances (Sprous and Salemme 2007). A comprehensive list of food/flavour related database is presented in Tables 13 and 14 depicts a list of databases and organizations (commercial as well as country specific) for agrochemicals related to plant protection and pest management along with issue of risk management and risk assessment. Still these databases are very small compared to drug discovery compound libraries in the industry. Recent initiatives requiring greater use of in silico technologies have called for transparency and development of appropriate database information that is available to the public at no cost. Therefore, the research community should take initiatives to develop more databases for public and administration uses in the fields of agrochemicals and food sciences.

Table 13 A comprehensive list of food/flavour related database (few database are consisting of pharmaceuticals and chemicals also)

Database	Content	Size (Approx.)	Web address
Flavor and Extract Manufacturers Association/Generally Recognized As Safe (FEMA/GRAS)	Flavors	2244	http://www.femaflavor.org
AnalytiCon	Natural products	2449	http://www.ac-discovery.com/
Specs NP	Natural products	467	
TCM	Natural products	32357	http://www.tcmpage.com/
SuperScent	Flavors and fragrances	2116/1591 M	http://bioinf-applied.charite.de/superscent/
Research Institute for Fragrance Materials/Fragrance and Flavor Database (RIFM/FEMA)	Flavors and fragrances	5100	http://www.rifm.org/index.php
International Organization of the Flavor Industry (IOFI)	Flavors	2800	http://www.iofi.org
Everything Added to Food in the United States (EAFUS) maintained by the US FDA Center for Food Safety and Applied Nutrition (CFSAN)	Substances directly added to food	2000	http://www.accessdata.fda.gov/scripts/fcn/fcnNavigation.cfm?filter=&sortColumn=&rpt=eafusListing&displayAll=false#1
Food Chemicals Codex (FCC)	Monographs of food-grade chemicals, processing aids, flavoring agents, vitamins, and functional food ingredients	1200	http://www.usp.org/food-ingredients/food-chemicals-codex
Flavor-Base Database of Flavoring Materials and Food Additives	Flavor, regulatory, toxicological, and related data relevant to the flavor, food, beverage, and tobacco industries	–	http://www.leffingwell.com/flavbase.htm
Volatile Compounds in Food Database (VCF)	Volatile compounds	8000	http://www.vcf-online.nl/VcfHome.cfm

(continued)

Table 13 (continued)

Database	Content	Size (Approx.)	Web address
The (Complete) Database of Essential Oils (ESO)	Published by the Boelens Aroma Chemical Information Service (BACIS). Essential oils including in some cases multiple samples of the same oil from different sources, or having different countries of growing origin	4100	http://www.leffingwell.com/baciseso.htm
Flavor and Fragrance Materials (FEM)	Information collected from a variety of sources, including flavor and fragrance suppliers, industry and government organizations, as well as related texts	–	http://dir.perfumerflavorist.com
Flavornet	Compilation of aroma compounds	730	http://www.flavornet.org

Table 14 A comprehensive lists of database and organization of agrochemicals related to plant protection and pest management

Database/Organization	Content	Web address
Chemdatas	Agrochemical Database contains 7749 Products including 1590 Agrochemical Products, 1042 Agrochemical Intermediate and 5000 other Chemical Raw Materials	http://www.chemdatas.com/Chemdatas/AdAgrochemE.aspx
Croponosis	The Agrochemical Products Database (APD) provides comprehensive commercial and technical data on almost 600 of the most important active ingredients in the agrochemicals market	http://www.croponosis.com/marketing/products/databases
Agrochemicals IUPAC	Contains information about Biopesticides, Codex Alimentarius, Disposal and Storage of Pesticides, History of Pesticide Use, Integrated Pest Management, Introduction to Pesticide Profiles, Obsolete Pesticides, Pesticide Formulations, Pesticide Resistance, Pesticide Specifications, Pesticides and	http://agrochemicals.iupac.org/

(continued)

Table 14 (continued)

Database/Organization	Content	Web address
	Minor Crops, Prior Informed Consent, Safe Use of Pesticides, Spray Drift and its Mitigation, Maximum Residue Levels (MRLs), Pesticide Residues in Water, Residue Analytical Methods, Residue Studies—Conduct of Trials, Risk Assessment	
Agricultural Chemical Usage Database	Initiative by Department of Environment, Government of Australia. Database contains certain information on the agricultural chemicals used in Australia from 1997 to 2006 by broadacre farmers. The database allows governments, chemical users and the community to view trends in the usage of chemicals over an entire decade	https://www.environment.gov.au/protection/chemicals-management/agricultural-chemical-usage-database
Homologa™, The Global Crop Protection Database	Gives access to information about: Approved plant protection products, Registration-Status of active ingredients on EU-level, Actual status of registration, Approved parallel imports, MRLs in foodstuffs, Export-/import statistics of food	http://www.homologa-new.com/pls/apex/f?p=550:1:0::::
Codex Pesticides Residues in Food Online Database	User can obtain information on Codex MRLs and Codex Extraneous Maximum Residue Limits (EMRLs) for pesticide/commodity combinations. Details of commodities are found in the Codex Classification of Foods and Animal Feeds	http://www.fao.org/fao-who-codexalimentarius/standards/pesticide-mrls/en/
EU Pesticides database	Follows Activesubstances-Regulation (EC) No 1107/2009 and Pesticides EU-MRLs-Regulation (EC) No 396/2005	http://ec.europa.eu/food/plant/pesticides/eu-pesticides-database/public/?event=homepage&language=EN
National Agricultural Statistics Service (NASS)	The NASS Agricultural Chemical Use Program is United States Department of Agriculture (USDA) official source of statistics about on-farm chemical use and pest management practices. Since 1990, NASS has surveyed U.S. farmers to collect information on the chemical ingredients they apply to agricultural commodities through fertilizers and pesticides	https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Chemical_Use/

(continued)

Table 14 (continued)

Database/Organization	Content	Web address
Phillips McDougall	Phillips McDougall offer a number of products and services aimed at providing detailed analysis of the agrochemical and seed industries, including AgriService, Seed Service and Agreworld. In addition to this the company undertakes consultancy studies on a confidential single client basis	https://www.phillipsmcdougall.com/home.asp
Greenbook	Worked directly with chemical plant protection manufacturers to compile a versatile database of plant protection products	http://www.greenbook.net/
The Japan Food Chemical Research Foundation	Contribute to the safety of food and the maintenance and enhancement of the health of people through various activities. The Foundation supports research activities aimed at developing safety evaluation methodologies for food additives, and at reducing the use of food additives along with agrochemicals residue in food	http://www.ffcr.or.jp/zaidan/FFCRHOME.nsf/pages/agri-chem
AgNIC (Agricultural Network Information Center)	Directory of quality agriculture-related databases, datasets, and information systems	https://www.agnic.org/
E-Answers	Extension and outreach publications produced by land grant universities. Information about agriculture, forestry, fishing, family/consumer issues, lawn and garden	http://adec.edu/
EXtensionTOXicologyNETwork (EXTOXNET)	Supplies toxicology information and fact sheets, Pesticide Information Profiles (PIPs), and Toxicology Information Briefs (TIBs)	http://extoxnet.orst.edu/
National Ag Safety Database	Supplies comprehensive information regarding agricultural safety and health education. Database includes state publications, OSHA and EPA standards	http://www.cdc.gov/nasd/
National Pesticide Information Center (NPIC)	NPIC provides objective, science-based information about pesticides and pesticide-related topics to enable people to make informed decisions about pesticides and their use. NPIC is a cooperative agreement between Oregon State University and the U.S. Environmental Protection Agency (EPA)	http://npic.orst.edu/about.html

(continued)

Table 14 (continued)

Database/Organization	Content	Web address
TOXNET	Searching databases on toxicology, hazardous chemicals, environmental health, and toxic releases including Agrochemicals	https://toxnet.nlm.nih.gov/
CDMS Label Database	Search engine for agrochemicals by manufacturer	http://www.cdms.net/Label-Database
Agricultural pests	Information about managing pests, including University of California's official guidelines for monitoring pests and using pesticides and nonpesticide alternatives for managing insect, mite, nematode, weed, and disease pests	http://ipm.ucanr.edu/PMG/crops-agriculture.html
EPA Pesticides Databases	Database resources about environmental effects, environmental fate, health effects, and regulatory information	https://www.epa.gov/pesticides
National Pesticide Information Retrieval System (NPIRS)	Search for federally active pesticide products using one of the following methods: EPA Registration Number, Product Name, Company Name or Active Ingredient	http://ppis.ceris.purdue.edu/
PAN Pesticide Database	One-stop location for toxicity and regulatory information for pesticides, insecticides, herbicides	http://www.pesticideinfo.org/
USDA Plant Health Portal	USDA's clearinghouse on plant diseases, pest management, weeds management and plant health research	http://www.usda.gov/wps/portal/usda/usdahome?navid=PLANT_HEALTH&navtype=RT&parentnav=TOPICS
Pesticide Data Program (PDP)	The PDP is a national pesticide residue monitoring program and produces the most comprehensive pesticide residue database in the U. S.	https://www.ams.usda.gov/datasets/pdp
European and Mediterranean Plant Protection Organization (EPPO)	Under the International Plant Protection Convention (IPPC), EPPO is the regional plant protection organization (RPPO) within the European and Mediterranean region. EPPO Global Database is web-based database which has the objective to gather all pest-specific information that has been produced by EPPO	https://www.eppo.int/
PHYTOWEB	Official information on plant protection products registered in Belgium	http://fytoweb.be/en

(continued)

Table 14 (continued)

Database/Organization	Content	Web address
AGES database	List of authorised/approved plant protection products in Austria	http://pmg.ages.at/pls/psmlfrz/pmgweb4\$.Startup
Positivlisten	Database on plant protection products of Denmark	https://www.middeldatabasen.dk/positiveList.asp
Agricultural Pesticides Committee (APC)	Provide up-to-date approved information on pesticides use, decrease the misuse of unapproved products, increase transparency with all stakeholders using pesticides, ensure the safe use of pesticides for people and the environment in Egypt. The APC supports the crucial necessity of establishing the “Egyptian Organization of Pesticide Management; EOPM” in Egypt; that may undertake such roles of the US-EPA	http://www.apc.gov.eg/en/default.aspx
Plant Protection Product Register database (Tukes—Finnish Safety and Chemicals Agency)	Contains key information about plant protection products authorised in Finland. In addition to the authorised use, the register contains usage restrictions for each product. The label texts including the classification of the product and instructions for use can also be found in the register	http://tukes.fi/en/Branches/Chemicals-biocides-plant-protection-products/Plant-protection-products/Authorised-products/Plant-Protection-Product-Register/
E-phy	The catalog of plant protection products and their uses, fertilizers and growing media allowed in France	https://ephy.anses.fr/
AGRITOX	Database maintained by ANSES on physical and chemical properties, toxicology and ecotoxicology of plant protection substances registered in France	http://www.agritox.anses.fr/
Union of Industries of Plant Protection (UIPP)	French Crop Protection Association—datasheets on safety aspects of commercial products	http://www.uipp.org/
BVL database	The Federal Office of Consumer Protection and Food Safety was founded in 2002 in Germany. The BVL is an independent higher federal authority within the Federal Ministry of Food and Agriculture	http://www.bvl.bund.de/DE/Home/homepage_node.html
Ministry of rural development and food-Greece	List of authorized plant protection products	http://www.minagric.gr/syspest/syspest_menu_eng.aspx
IOBC Pesticide Side Effect Database	The database on selectivity of pesticides has been prepared jointly by the International Organization for Biological and Integrated Control West	http://www.iobc-wprs.org/index.html

(continued)

Table 14 (continued)

Database/Organization	Content	Web address
	Palearctic Regional Section (IOBC-WPRS) Working Group on "Pesticides and Beneficial Organisms" and the Commission "Guidelines for Integrated Production" to assist organizations and growers in the choice of pesticides	
Pesticides Registration & Control Division (PRCD)	The Database founded by Department of Agriculture and Food, Government of Ireland, contains details of registered plant protection products and is updated on a regular basis. It can be searched by Product Name, Active Substance or Function/Crop	http://www.pcs.agriculture.gov.ie/products/
Plant Protection and Inspection Services (PPIS)	The database contains information on safe and proper usage of the approximately 900 pesticides registered in Israel. Database developed by Ministry of Agriculture and Rural Development of Israel	http://www.hadbara.moag.gov.il/hadbara/english/
State Plant Protection Service (SPPS)	SPPS is a public administration body to provide the national phytosanitary safety by taking effective monitoring measures to protect the country from dangerous plant diseases and pests, and provide plant and plant product exports. SPPS built on 17 December 1998, adopted by the Plant Protection Act, and the Ministry of Agriculture of Latvia	http://www.vaad.gov.lv/sakums/par-mums.aspx
ASTA	List of registered plant protection products of Luxembourg	https://saturn.etat.lu/tapes/
Index Phyto Sanitary	Contains data on agricultural pesticides registered product Morocco regardless whether marketed or not	http://eservice.onssa.gov.ma/IndPesticide.aspx
CTGB	Board for the Authorisation of Plant Protection Products (specifically, pesticides) and Biocides of the Netherlands	http://www.ctgb.nl/en/pesticides-database
Mattilsynet	Database on registered plant protection products of Norway	http://www.mattilsynet.no/plantevermidler/godk.asp?sortering=preparat&preparat=Alle&sprak=In+English
Poland plant protection database	Online database on plant protection products of Poland by Ministry of Agriculture and Rural Development	http://www.minrol.gov.pl/eng/Ministry/Online-database-on-plant-protection-products

(continued)

Table 14 (continued)

Database/Organization	Content	Web address
Health and Safety Executive databases (HSE)	Guidance on authorisation for pesticides used in Agriculture, Horticulture or the Home Garden in United Kingdom. Guidance on how to use these products safely and information about controls over pesticide residues in food	http://www.hse.gov.uk/pesticides/
UKSUP	Authorized Plant Protection Products of Slovakia	http://www.uksup.sk/index.php?start&t=orp-pripravky-na-ochranu-rastlin-registre-a-zoznamy&t2=
KEMI	The Pesticides Register contains information on more than 3,000 Authorised and Previously Authorized pesticide products in Sweden	http://webapps.kemi.se/BkmRegistret/Kemi.Spider.Web.External/

10 Expert Systems

Expert systems allow for the direct entry of a structure into the software followed by the calculation or prediction of the response without the requirement to compute descriptors and re-perform the modeling process. This makes expert systems a more convenient option for quick prediction of activity/response over traditional QSARs. Expert systems have been frequently employed by regulatory agencies, academia and industries worldwide for more efficient and fast prediction. Multiple mechanisms can lead to the same activity/response and this complexity requires the availability of predictive tools that are able to distinguish multiple regions in the activity space. This need has led to the development of so-called expert systems, which try to cover broader structural and activity regions in comparison with the local models. There are a number of commercial or freely available toxicity prediction software packages. They have advantages over the use of traditional QSARs. A representative list of free and commercially available QSAR expert systems has been portrayed in Table 15.

11 Conclusion

This chapter presents the current knowledge related to successful QSARs application in the field of food and agriculture sciences along with the information on how one can tactfully utilize this technique for exploring better and efficient food products, food supplements as well as agrochemicals and phytochemicals. The results of a number of studies are analyzed and discussed to provide a solid rationale for continuing efforts to improve QSAR models in the food and

Table 15 A representative list of various freely available and commercial software and online resources in QSAR studies

Source	Name of the program/tool	Executable operations	Web address
US EPA	Representative names of freely available QSAR expert systems/tools		
	ECOSAR	Allows QSAR based prediction of toxicity potential of chemicals to fish, invertebrates, and aquatic plants along with a limited number of methods for terrestrial and marine organisms	http://www.epa.gov/tsca-screeningtools/ecological-structure-activityrelationships-ecosar-predictive-model
	OncoLogic™	Allows determination of cancer potential of chemicals using mechanism based SAR analysis in conjugation with a decision tree based formalism	http://www.epa.gov/tsca-screeningtools/oncologictm-computer-system-evaluate-carcinogenic-potential-chemicals
	Non-Cancer Health Assessment	Allows identification of health effects data using various public databases	http://www.epa.gov/tsca-screeningtools/non-cancer-screening-approacheshhealth-effects
	AIM	Allows analogue based searching of data to facilitate assessment of chemical hazard or read-across operation	http://www.epa.gov/tsca-screeningtools/analog-identification-methodologyaim-tool
	ChemACE	Allows identification of analogous chemicals for potential read across using structural diversity based recognition methodology	http://www.epa.gov/tsca-screeningtools/chemical-assessment-clusteringengine-chemace
	EPI Suite™	An assemblage of tools performing prediction of physical/chemical properties and environmental fate attributes <ul style="list-style-type: none"> ● KOWWIN™: Estimating log octanol-water partition coefficient. ● AOPWIN™: Estimating gas-phase reaction rate of chemicals with mostly prevalent atmospheric oxidants, hydroxyl radicals, ozone radicals etc. ● HENRYWIN™: Estimating Henry's Law constant (air/water partition coefficient) employing group contribution and bond contribution formalisms ● MPBPWIN™: Melting point, boiling point, and vapor pressure of organic chemicals are estimated using a combination of techniques. Included is the subcooled liquid vapor pressure, which is the vapor pressure a solid would have if it were liquid at room temperature. It is important in fate modeling ● BLOWIN™: Determining aerobic and anaerobic biodegradability of organic chemicals using different predictive models ● BioHCwin: Estimating biodegradation half-life for hydrocarbons ● KOCWIN™: Determining the organic carbon-normalized sorption coefficient for soil and sediment using multiple methods ● BCFBAF™: Estimating fish bio-concentration factor and its logarithm employing multiple methods ● HYDROWIN™: Computing aqueous hydrolysis rate constants and half-lives of esters, carbamates, epoxides, halomethanes, selected alkyl halides, and phosphorus esters 	http://www.epa.gov/tsca-screeningtools/epi-suite-estimation-program-interface

(continued)

Table 15 (continued)

Source	Name of the program/tool	Executable operations	Web address
		<p>Executable operations</p> <ul style="list-style-type: none"> • AEROWIN™: Estimating the fraction of airborne substance sorbed to airborne particulates (ϕ, phi) using different methods • WVOWIN™: Estimating the rate of volatilization of chemicals followed by determination of half-lives • LEV3EPI™: This is a multimedia (level III) fugacity model and allows estimation of partitioning behavior of chemicals into air, soil, sediment, and water compartments under steady state conditions • Related tools using input from the above mentioned programs include WSKOWWIN™, WATERNT™, KOAWIN, STPWIN™ <p>Allows estimation of environmental releases and workplace exposure of chemicals which are manufactured and used in industrial and commercial houses</p>	<p>http://www.epa.gov/tscascreeningtools/chemsteer-chemical-screening-tool-exposures-and-environmental-releases</p>
	ChemSTEER	Allows estimation of environmental releases and workplace exposure of chemicals which are manufactured and used in industrial and commercial houses	http://www.epa.gov/tscascreeningtools/chemsteer-chemical-screening-tool-exposures-and-environmental-releases
	E-FAST	Allows estimation of chemical released to air, landfills, surface water, and consumer products	http://www.epa.gov/tscascreening-tools/efast-exposure-and-fate-assessmentscreening-tool-version-2014
	ReachScan	A tool under development to determine chemical concentrations in surface water at the stream segments downstream from industrial facilities	http://www.epa.gov/tscascreeningtools/reachscan-exposure-assessment-model
	TEST	Allows QSAR model based prediction of various physical properties (boiling point, viscosity, density, flash point etc.), toxicity endpoints (daphnid, fathead minnow, tetrahymena, rat toxicity etc.) and bio-concentration factor	http://www.epa.gov/chemicalresearch/toxicity-estimation-software-tooltest
QSPR-Thesaurus	QSPR-Thesaurus	Allows access to stored data as well as prediction of various property and toxicity responses of chemicals	http://qspr-thesaurus.eu/home/show.do
OCHEM	OCHEM	Allows model development as well as model based prediction of various property and toxicity endpoints	https://ochem.eu/home/show.do
Organic Chemistry Portal	OSIRIS property Explorer	Allows computation of various properties using a color coding system to identify drug like behavior	http://www.organic-chemistry.org/prog/peo/
VCCLAB	VCCLAB	Allows computation of various physicochemical and structural variables along with options for performing chemometric operations	http://www.vcclab.org/
Heimholtz Centre for Environmental Research, UFZ	ChemProp	Allows computation of various environmentally relevant physical-chemical properties including partitioning, degradation, ecotoxicity, environmental fate e.g., BCF, BAF etc.	http://www.ufz.de/index.php?en=6738

(continued)

Table 15 (continued)

Source	Name of the program/tool	Executable operations	Web address
Opentox project European Commission	ToxPredict	Allows interoperable toxicological assessment of chemicals	http://www.opentox.org/
PaDEL	PaDELDDPredictor	Allows computation of pharmacodynamics, pharmacokinetics and toxicological properties	http://www.yapcwsoft.com/dd/padelddpredictor/
OECD	OECD QSAR Toolbox	Allows assessment of ecotoxicological hazard of chemicals using QSAR model based prediction and aids in filling data gaps	http://www.oecd.org/chemicalsafety/riskassessment/theoecdqsartoolbox.htm
VEGA	VEGA	Allows QSAR model based prediction of various environmental, ecotoxicological and toxicological endpoints as well as various chemometric operations like molecular dynamics analysis, virtual screening, molecular mechanics based calculations, conformational search, docking etc. using two different platforms	http://www.vega-qsar.eu/
toxRead	toxRead	Allows assessment of mutagenicity and read across evaluation of chemicals	http://www.tox.gate.eu/index.php
European Commission	ToxTree	Allows prediction of toxicity viz. mutagenicity, skin irritation, eye irritation, biodegradation etc., and categorization of chemicals into different schemes employing decision tree formalism	https://eur-lexvam.jrc.ec.europa.eu/laboratoriesresearch/predictive_toxicology/qsar_tools/toxtree
	Toxmatch	Allows read-across and categorization of chemical hazard. Toxicity endpoints include skin irritation, skin sensitisation, bioconcentration factor and aquatic toxicity	https://eur-lexvam.jrc.ec.europa.eu/laboratoriesresearch/predictive_toxicology/qsar_tools/toxmatch
	DART	Decision Analysis by Ranking Techniques, allows ranking of chemicals based on their hazardous potential	https://eur-lexvam.jrc.ec.europa.eu/laboratoriesresearch/predictive_toxicology/qsar_tools/DART
	Stat4tox	Software for the Statistical Evaluation of in vitro Assays, allows statistical analysis of toxicological data as well as dose-response modelling for the estimation of EC50 value of chemicals	https://eur-lexvam.jrc.ec.europa.eu/laboratoriesresearch/predictive_toxicology/qsar_tools/stat4tox
	METIS	Metabolic Information Input System allows inspection of metabolism and degradation reaction data. Uses the Chemical Reactivity and Fate Tool (CRAFT) database	https://eur-lexvam.jrc.ec.europa.eu/laboratoriesresearch/predictive_toxicology/qsar_tools/METIS
	JRC QSAR Model Database	A well-documented repository of QSAR models to facilitate toxicological assessment of chemicals following REACH guideline	https://eur-lexvam.jrc.ec.europa.eu/laboratoriesresearch/predictive_toxicology/qsar_tools/QRf
	ToxBank Project	Helps in assessment of chemical hazard with its toxicity data repository coupled with a cross-clustered integrated data analysis	http://toxbank.net/

(continued)

Table 15 (continued)

Source	Name of the program/tool	Executable operations	Web address
Laboratory of Mathematical Chemistry-OASIS	QSAR TOOLBOX	Performs chemical categorization using QSAR modeling, read across, trend analysis to aid data gap filling and hazardous assessment of chemicals	http://oasis-lmc.org/
	CATALOGIC	Allows assessment of environmental fate and ecotoxicity of chemicals	
	TIMES	Allows prediction of toxicity of chemical metabolites	
	POPs	Allows assessment of PBT profile of chemicals	
Insilico toxicology GmbH	Lazar	Lazy Structure–Activity Relationships allows prediction of toxicity to fish lethality, carcinogenicity towards various endpoints like hamster, mouse, rat etc., mutagenicity and repeated dose toxicity	http://hazar.in-silico.de/predict
Delta Group and Virtua Drug	Althotas	Allows toxicokinetics assessment by in silico prediction of biological properties of given proteins using support vector machine and de novo protein binding formalism	http://www.althotas.com/
The European Chemical Industry Council	LRI Toolbox	Various tools allowing chemical categorization by providing a database containing significant amount of chemical data characterizing environmental and human health. AMBIT, ART, BiotS, Busy, IndusChemFate, OLIMPIC CRAFT etc. are some of the human health models while ADEPT, GEMCO, GREAT-ER, TERRACE etc. represent environmental models	http://cefic-lri.org/
The Canadian Centre for Environmental Modelling and Chemistry	BASL4	Biosolids-Amended Soil Level IV allows prediction of chemical fate introduced to soil along with contaminated biosolids amendment. The processes of chemical degradation, volatilization, leaching, diffusion, sorbed phase transport due to bioturbation, and the degradation of the organic matter (OM) present in the soil	http://www.trentu.ca/academic/aminss/envmodel/models/BASL4110.html
Representative names of commercially available QSAR expert systems/tools			
CambridgeSoft	ChemOffice	Allows prediction of various properties e.g., pKa, LogP, LogS, melting point, boiling point, molar refraction, heat of formation etc.	http://www.cambridgesoft.com/
Technical University of Denmark	ICAS-Toolboxes	ProPred: For estimating pure component properties employing different methods	http://www.capec.kit.dtu.dk/Software/ICASand-its-Tools/ICAS-Toolboxes
		PDS: For investigating ternary mixtures with respect to distillation based separation	
ACD/Labs	ACD/Percepta	Allows model based prediction of various physicochemical properties, ADME measures, and toxicity endpoints	http://www.acdlabs.com/
DassaultSystèmes	BIOVIA	Allows property prediction, chemometric analysis and safety data management of chemicals using various modules like CISPro, QSAR Workbench, Discovery Studio, LIMS, TOPKAT etc.	http://accelrys.com/products/

(continued)

Table 15 (continued)

Source	Name of the program/tool	Executable operations	Web address
Simulations Plus Inc	ADMET Predictor™	Allows predictive modeling of ADMET properties	http://www.simulations-plus.com/
	GastroPlus™	Allows PBPK modeling and simulation	
	MedChem Studio™	Allows various in silicoligand based operations facilitating identification and optimization of lead	
Fujitsu	SCIGRESS and ADF	Runs various chemometric operations	http://www.fqs.jp/
	ADMEWORKS ModelBuilder	Allows developing predictive QSAR models	
	ADMEWORKS Predictor	Allows in silico virtual screening and ADMET property evaluation	
TimTec Inc.	ChemDBsoft	Several packages namely MOLPRO, SLIPPER, DISCON, HYBOT allows computation of drug like properties viz. solubility, permeability, pKa, lipophilicity along with various chemometric operations	http://www.chemdbsoft.com/
Bioinformatics and Molecular Design Research Centre	PreADMET	Allows prediction of drug likeness, ADME properties and toxicity (mutagenicity, carcinogenicity) of chemicals	http://preadmet.bmdrc.kr/
Schrödinger	Schrödinger	Allows property prediction along with various in silico operations using various products (modules) e.g., QikProp, EpiK, Maestro, MacroModel, BioLuminate, SARvision, Glide, MacroModel etc.	http://www.schrodinger.com/products/
Molecular discovery	MoKa	Computation of pKa values	http://www.moldiscovery.com/
	VolSurf+	Aids ADME prediction by employing molecular interaction field	
	Pentacle, GRID, SHOP, FLAP, MetaDesign	Various fingerprint, alignment etc. based molecular modeling and QSAR operations	
GeneXplain	PASS	Performs model based prediction of biological activity, biochemical mechanism, toxicity, metabolism, gene regulation expression and transporter related attributes	http://www.genexplain.com/pass
MolCode Ltd.	QSAR Service™ (MolCode)	Allows QSAR model based prediction of REACH and OECD guidelines compliant toxicity endpoints namely skin sensitisation, skin irritation, eye irritation, mutagenicity, inhalation toxicity, developmental toxicity, toxicity to daphnids, algae, fish and birds, biodegradation etc.	http://www.molcode.com/

(continued)

Table 15 (continued)

Source	Name of the program/tool	Executable operations	Web address
Multicase	CASE Ultra	Allows prediction of toxicity and bioactivity following ICH M7 guidelines	http://www.multicase.com/
Leadscope	Leadscope	Uses various expert alerts and statistical models allowing prediction of various toxicity endpoints and adverse effects on systems like hepatobiliary, urinary tract, cardiovascular etc.	http://www.leadscope.com/
CompuDrug	HazardExpert	Allows computation of different toxicity outcomes involving oncogenicity, mutagenicity, teratogenicity, membrane irritation, sensitivity, immunotoxicity, and neurotoxicity as well as bioavailability and bioaccumulation potential of chemicals	http://www.compuDrug.com/hazardexpertpro
Lhasa Ltd.	Derek Nexus	Provides prediction for toxicity endpoints namely carcinogenicity, mutagenicity, genotoxicity, skin sensitization, teratogenicity, irritation, respiratory sensitization and reproductive toxicity	http://www.lhasalimited.org/derek_nexus/
TerraBase Inc.	TerraQSAR™ and TerraTox™	Allows toxicity estimation and environmental risk assessment of pharmaceuticals, pesticides, and waste management. Toxicity endpoints include carcinogenicity, skin irritation, toxicity to rat, mouse, daphnids, fathead minnow, estrogen binding affinity and partition coefficient of chemicals	http://www.terrabase-inc.com/
BioByte	Bio-Loom, CQSAR	Allows QSAR model based predictions of partition coefficient (ClogP) and toxicity endpoints	http://www.biobyte.com/bb/prod/cqsarad.html
Thomson Reuters	MetaDrug™	Contains manually curated biological data on small molecules and allows QSAR model based prediction of toxicity, ADME properties and therapeutic activity of compounds	http://thomsonreuters.com/en/productservices/pharma-lifesciences/pharmaceuticalresearch/metadug.html
Cyprotex	chemTox	Allows prediction of toxicity without using any in vitro physicochemical or toxicity information. Endpoints include Ames mutagenicity and oral/iv administration based rat acute toxicity (LD ₅₀)	http://www.cyprotex.com/insilico/physiological_modelling/chemtox

agriculture sciences. However, lack of sufficient data related to a particular class of molecules has hindered the allocation of computational modeling methods to some extent. Moreover, among the significant number of QSAR models developed, majority have been concentrated on the development of models for specific class of derivatives.

Considering future perspectives, with the increasing acceptance that chemical diversity of plant products is well suited to provide core scaffolds for future drugs, there is an increasing use of novel plant products and chemical libraries based on phytochemicals in drug discovery programs. Nevertheless, only a limited number of *in silico* models have been reported in the literature so far based on phytochemicals. The world may truly benefit from the wealth of knowledge of traditional plant medicine on successful integration of ethnopharmacology and *in silico* approaches.

There is no doubt that QSAR has come a long way from its inception days in the form of classical Hansch and Free-Wilson approaches. It has gradually evolved with the use of newer descriptors, rapid advances in computer power, application of diverse chemometric tools, employment of rigorous validation tests, and integration with receptor structure information. QSAR has now emerged as a distinct scientific discipline in its own right. A good practice of QSAR modeling through usage of OECD-recommended guidelines can develop good predictive models with demonstrated practical applications in diverse chemical—biological areas as applied in the fields of food and agriculture, which may further strengthen its acceptability to the scientific community. There are many novel molecules tested for antioxidant activity that are in the pipeline of pre-clinical and clinical trials from the application of QSAR approach. It is also true for agrochemicals. In Fig. 31 we depicted already commercialized agrochemicals which have come to the market with the assist of classical QSAR. We expect that applications of chemoinformatic methods will intensify in future in all areas where new chemicals are being developed and tested.

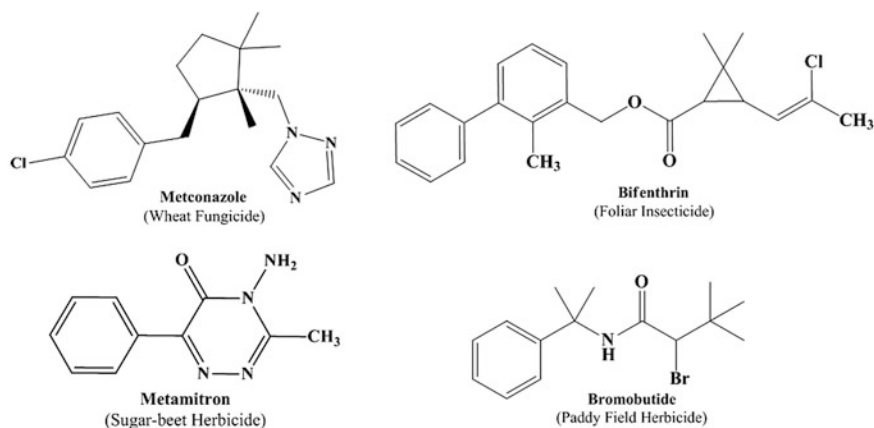


Fig. 31 Important examples of commercialized agrochemicals with the aid of classical QSAR

Acknowledgements S.K. and J.L. thank the National Science Foundation (NSF/CREST HRD-1547754, and EPSCoR (Award#: 362492-190200-01\NSFEP5-0903787) for financial support. K.R. thanks the UGC for financial assistance under the UPEII programme.

References

- AGES. (2010). Impact of metabolic and degradation processes on the toxicological properties of residues of pesticides in food commodities. Report from the Austrian Agency for Health and Food Safety (AGES) to the European Food Safety Authority (EFSA). Retrieved August 18, 2016, from <http://www.efsa.europa.eu/en/scdocs/scdoc/49e.htm>.
- Anczewicz, J., Migliavacca, E., Carrupt, P.-A., Testa, B., Brée, F., Zini, R., ..., Le Ridant, A. (1998). Structure–Property relationships of Trimetazidine derivatives and model compounds as potential Antioxidants. *Free Radical Biology and Medicine*, 25(1), 113–120.
- Arvidson, K. B., Valerio, L. G., Diaz, M., & Chanderbhan, R. F. (2008). In Silico Toxicological screening of natural products. *Toxicology Mechanisms and Methods*, 18(2–3), 229–242.
- Arvidson, K. B., Chanderbhan, R., Muldoon-Jacobs, K., Mayer, J., & Ogungbesan, A. (2010). Regulatory use of computational toxicology tools and databases at the United States food and drug administration’s office of food additive safety. *Expert Opinion on Drug Metabolism and Toxicology*, 6(7), 793–796.
- Auer, C. M., Nabholz, J. V., & Baetcke, K. P. (1990). Mode of action and the assessment of chemical hazards in the presence of limited data: Use of structure-activity relationships (SAR) under TSCA, section 5. *Environmental Health Perspectives*, 87, 183–197.
- Bailey, A. B., Chanderbhan, R., Collazo-Braier, N., Cheeseman, M. A., & Twaroski, M. L. (2005). The use of structure–activity relationship analysis in the food contact notification program. *Regulatory Toxicology and Pharmacology*, 42(2), 225–235.
- Basant, N., Gupta, S., & Singh, K. P. (2015). Predicting Toxicities of diverse chemical pesticides in multiple avian species using tree-based QSAR approaches for regulatory purposes. *Journal of Chemical Information and Modeling*, 55(7), 1337–1348.
- Beattie, G. A. (2006). Phytobacteria with a commensalistic association with plants. In S. S. Gnanamanickam (Ed.), *Plant associated bacteria* (pp. 30–35). Netherlands: Springer.
- Beltrán, H. I., Damian-Zea, C., Hernández-Ortega, S., Nieto-Camacho, A., & Ramírez-Apan, M. T. (2007). Synthesis and characterization of di-phenyl-tinIV-salicyliden-ortho-aminophenols: Analysis of in vitro antitumor/antioxidant activities and molecular structures. *Journal of Inorganic Biochemistry*, 101(7), 1070–1085.
- Benfenati, E. (2012). *The e-Book on QSAR and REACH: Theory, Guidance and Applications*. Retrieved August 18, 2016, from http://ebook.insilico.eu/insilico-ebook-orchestra-benfenati-ed1_rev-June2013.pdf.
- Bermúdez-Saldaña, J. M., & Cronin, M. T. (2006). Quantitative structure–activity relationships for the toxicity of organophosphorus and carbamate pesticides to the rainbow trout *Oncorhynchus mykiss*. *Pest Management Science*, 62(9), 819–831.
- Beydon, D., Payan, J.-P., Ferrari, E., & Grandclaude, M.-C. (2014). Percutaneous absorption of herbicides derived from 2, 4-dichlorophenoxyacid: Structure–activity relationship. *Toxicology in Vitro*, 28(5), 1066–1074.
- Bitencourt, M., & Freitas, M. P. (2008). MIA-QSAR evaluation of a series of sulfonylurea herbicides. *Pest Management Science*, 64(8), 800–807.
- Bredsdorff, L., Wedeby, E. B., Nikolov, N. G., Hallas-Møller, T., & Pilegaard, K. (2015). Raspberry ketone in food supplements—High intake, few toxicity data—A cause for safety concern? *Regulatory Toxicology and Pharmacology*, 73(1), 196–200.
- Buchbauer, G., Klein, C. T., Wailzer, B., & Wolschann, P. (2000). Threshold-based Structure–Activity relationships of Pyrazines with bell-pepper flavor. *Journal of Agricultural and Food Chemistry*, 48(9), 4273–4278.

- Buchholtz, K. P. (1967). Report of the terminology committee of the Weed Science Society of America. *Weeds*, 15, 388–389.
- Calgarotto, A. K., Miotto, S., Honório, K. M., da Silva, A. B. F., Marangoni, S., Silva, J. L., ... da Silva, S. L. (2007). A multivariate study on flavonoid compounds scavenging the peroxy nitrite free radical. *Journal of Molecular Structure: THEOCHEM*, 808(1–3), 25–33.
- Cao, X., Xu, S., Li, X., Shen, X., Zhang, Q., Li, J., et al. (2012). N-Nitrourea derivatives as novel potential fungicides against *Rhizoctonia solani*: Synthesis, Antifungal activities, and 3D-QSAR. *Chemical Biology and Drug Design*, 80(1), 80–88.
- Casalegno, M., Sello, G., & Benfenati, E. (2006). Top-priority fragment QSAR approach in predicting pesticide aquatic toxicity. *Chemical Research in Toxicology*, 19(11), 1533–1539.
- Chana, A., Tromelin, A., Andriot, I., & Guichard, E. (2006). Flavor release from ι-carrageenan matrix: A quantitative structure–property relationships approach. *Journal of Agricultural and Food Chemistry*, 54(10), 3679–3685.
- Chen, B., Zhu, Z., Chen, M., Dong, W., & Li, Z. (2014). Three-dimensional quantitative structure–activity relationship study on antioxidant capacity of curcumin analogues. *Journal of Molecular Structure*, 1061, 134–139.
- Chen, Y., Xiao, H., Zheng, J., & Liang, G. (2015). Structure-thermodynamics-antioxidant activity relationships of selected natural phenolic acids and derivatives: An experimental and theoretical evaluation. *PLoS ONE*, 10(3), e0121276.
- Cheng, Z., Ren, J., Li, Y., Chang, W., & Chen, Z. (2002). Study on the multiple mechanisms underlying the reaction between hydroxyl radical and phenolic compounds by qualitative structure and activity relationship. *Bioorganic and Medicinal Chemistry*, 10(12), 4067–4073.
- CIOMS. (1985). International Guiding Principles for Biomedical Research Involving Animals. Retrieved August 18, 2016, from http://cioms.ch/publications/guidelines/1985_texts_of_guidelines.htm.
- Copping, L. G., & Hewitt, H. G. (1998a). *Fungicides, chemistry and mode of action of crop protection agents* (pp. 74–113). Cambridge: The Royal Society of Chemistry.
- Copping, L. G., & Hewitt, H. G. (1998b). *Insecticides, chemistry and mode of action of crop protection agents* (pp. 46–73). Cambridge: The Royal Society of Chemistry.
- CRD. (2010). Applicability of thresholds of toxicological concern in the dietary risk assessment of metabolites, degradation and reaction products of pesticides. Report from the UK Chemicals Regulation Directorate (CRD) to the European Food Safety Authority (EFSA). Retrieved August 18, 2016, from <http://www.efsa.europa.eu/en/scdocs/scdoc/44e.htm>.
- Dambolena, J. S., Zygadlo, J. A., & Rubinstein, H. R. (2011). Antifumonisin activity of natural phenolic compounds a structure–property–activity relationship study. *International Journal of Food Microbiology*, 145(1), 140–146.
- Dambolena, J. S., López, A. G., Meriles, J. M., Rubinstein, H. R., & Zygadlo, J. A. (2012). Inhibitory effect of 10 natural phenolic compounds on *Fusarium verticillioides*. A structure–property–activity relationship study. *Food Control*, 28(1), 163–170.
- Denisov, E. T., & Afanasev, I. B. (2005). *Oxidation and antioxidant in organic chemistry and biology*. Boca Raton: CRC Press.
- Díaz, G. A., & Delgado, E. J. (2009). Quantitative prediction of AHAS inhibition by pyrimidinylsalicylate based herbicides. *Pesticide Biochemistry and Physiology*, 95(1), 33–37.
- Du, H., Wang, J., Hu, Z., Yao, X., & Zhang, X. (2008). Prediction of Fungicidal activities of rice blast disease based on least-squares support vector machines and project pursuit regression. *Journal of Agricultural and Food Chemistry*, 56(22), 10785–10792.
- Durand, A.-C., Farce, A., Carato, P., Dilly, S., Yous, S., Berthelot, P., et al. (2007). Quantitative structure-activity relationships studies of antioxidant hexahydropyridoindoles and flavonoid derivatives. *Journal of Enzyme Inhibition and Medicinal Chemistry*, 22(5), 556–562.
- EC. (1991). Council Directive 91/414/EEC of 15 July 1991 concerning the placing of plant protection products on the market. Official Journal of the European Union, L 230/1 of 19.08.1991. Office for Official Publications of the European Communities (OPOCE), Luxembourg.

- EC. (2002). Regulation (EC) No 178/2008 of the European Parliament and of the Council of 28 January 2002 laying down the general principles and requirements of food law, establishing the European Food Safety Authority and laying down procedures in matters of food safety. *Official Journal L*, 031, 1–24.
- EC. (2005). Regulation (EC) No 396/2005 of the European Parliament and of the Council of 23 February 2005 on maximum residue levels of pesticides in or on food and feed of plant and animal origin and amending Council Directive 91/414/EEC. *Official Journal, L070*, 1–16.
- ECHA. (2008). Guidance on Information Requirements and Chemical Safety Assessment. Chapter R6. European Chemicals Agency, Helsinki, Finland. Retrieved August 18, 2016, from http://guidance.echa.europa.eu/docs/guidance_document/information_requirements_en.htm?time=1252064523#r6.
- El-Masri, H. A., Mumtaz, M. M., Choudhary, G., Cibulas, W., & De Rosa, C. T. (2002). Applications of computational toxicology methods at the agency for toxic substances and disease registry. *International Journal of Hygiene and Environmental Health*, 205(1–2), 63–69.
- Erzincan, P., Saçan, M. T., Yüce-Dursun, B., Daniş, Ö., Demir, S., Erdem, S. S., et al. (2015). QSAR models for antioxidant activity of new coumarin derivatives. *SAR and QSAR in Environmental Research*, 26(7–9), 721–737.
- Eskes, C., & Zuang, V. (2005). Alternative (non-animal) methods for cosmetics testing: current status and future prospects. A report prepared in the context of the 7th amendment of the cosmetics directive for establishing the timetable for phasing out animal testing. *ATLA*, 33 (suppl 1):19–20.
- EU. (2006). Regulation (EC) No 1907/2006 of The European Parliament and of the Council concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *Official Journal of the European Union, L396*, 1–843.
- Farkas, O., Jakus, J., & Héberger, K. (2004). Quantitative structure—Antioxidant activity relationships of Flavonoid compounds. *Molecules*, 9(12), 1079–1088.
- FDA. (2002). Guidance for industry: Carcinogenicity study protocol submissions. FDA Center for Drug Evaluation and Research (CDER).
- FDA. (2014). Consumers—Dietary supplements: what you need to know. Retrieved August 18, 2016, from <http://www.fda.gov/>.
- Fratamico, P.M., & Bayles, D.O. (2005). *Foodborne pathogens: Microbiology and molecular biology*. Caister Academic Press.
- Fratev, F., Piparo, E. L., Benfenati, E., & Mihaylova, E. (2007). Toxicity study of allelochemical-like pesticides by a combination of 3D-QSAR, docking, local binding energy (LBE) and GRID approaches. *SAR and QSAR in Environmental Research*, 18(7–8), 675–692.
- Fujio, M., McIver, R. T., & Taft, R. W. (1981). Effects of the acidities of phenols from specific substituent-solvent interactions. Inherent substituent parameters from gas-phase acidities. *Journal of the American Chemical Society*, 103(14), 4017–4029.
- Gu, Y., Majumder, K., & Wu, J. (2011). QSAR-aided in silico approach in evaluation of food proteins as precursors of ACE inhibitory peptides. *Food Research International*, 44(8), 2465–2474.
- Halliwell, B., & Gutteridge, J. M. C. (1990). The antioxidants of human extracellular fluids. *Archives of Biochemistry and Biophysics*, 280(1), 1–8.
- Hamadache, M., Benkortbi, O., Hanini, S., Amrane, A., Khaouane, L., & Si Moussa, C. (2016). A quantitative structure activity relationship for acute oral toxicity of pesticides on rats: Validation, domain of application and prediction. *Journal of Hazardous Materials*, 303, 28–40.
- Hansch, C., Maloney, P. P., Fujita, T., & Muir, R. M. (1962). Correlation of biological activity of Phenoxyacetic acids with Hammett Substituent constants and partition coefficients. *Nature*, 194(4824), 178–180.

- Hansch, C., Muir, R. M., Fujita, T., Maloney, P. P., Geiger, F., & Streich, M. (1963). The correlation of biological activity of plant growth regulators and Chloromycetin derivatives with Hammett constants and partition coefficients. *Journal of the American Chemical Society*, 85 (18), 2817–2824.
- Hartman, G.L., Bonde, M.R., Miles, M.R., Frederick, R.D. (2004). Variation of Phakopsora-chytrizi isolates on soybean. Proceedings. VII World Soybean Research Conference, pp. 440–446.
- Hasebe, A., Nishiwaki, H., Akiyama, K., Sugahara, T., Kishida, T., & Yamauchi, S. (2013). Quantitative Structure-Activity relationship analysis of Antifungal (+)-Dihydroguaiaretic acid using 7-Phenyl derivatives. *Journal of Agricultural and Food Chemistry*, 61(36), 8548–8555.
- He, Y., Niu, C., Wen, X., & Xi, Z. (2013). Biomacromolecular 3D-QSAR to decipher molecular herbicide resistance in acetohydroxyacid synthases. *Molecular Informatics*, 32(2), 139–144.
- He, D., Jian, W., Liu, X., Shen, H., & Song, S. (2015). Synthesis, biological evaluation, and structure-activity relationship study of novel Stilbene derivatives as potential Fungicidal agents. *Journal of Agricultural and Food Chemistry*, 63(5), 1370–1377.
- Helguera, A., Combes, R., Gonzalez, M., & Cordeiro, M. N. (2008). Applications of 2D Descriptors in drug design: A DRAGON tale. *Current Topics in Medicinal Chemistry*, 8(18), 1628–1655.
- Hengstler, J. G., Foth, H., Kahl, R., Kramer, P. J., Lilienblum, W., Schulz, T., et al. (2006). The REACH concept and its impact on toxicological sciences. *Toxicology*, 220, 232–239.
- Jacobs, A. (2005). Prediction of 2-year carcinogenicity study results for pharmaceutical products: How are we doing? *Toxicological Sciences*, 88(1), 18–23.
- Janse, D. (2005). Prevention and control of bacterial pathogens and diseases. *Phytopathology principles and practice* (pp. 149–174). Wallingford: CAB International.
- Jiang, D.-P., Zhu, C.-C., Shao, X.-S., Cheng, J.-G., & Li, Z. (2015). Bioactive conformation analysis of anthranilic diamide insecticides: DFT-based potential energy surface scanning and 3D-QSAR investigations. *Chinese Chemical Letters*, 26(6), 662–666.
- Jing, P., Zhao, S., Ruan, S., Sui, Z., Chen, L., Jiang, L., et al. (2014). Quantitative studies on structure–ORAC relationships of anthocyanins from eggplant and radish using 3D-QSAR. *Food Chemistry*, 145, 365–371.
- JRC. (2010). Applicability of QSAR analysis to the evaluation of the toxicological relevance of metabolites and degradates of pesticide active substances for dietary risk assessment. Report from the European Commission’s Joint Research Centre (JRC) to the European Food Safety Authority (EFSA). Retrieved August 18, 2016, from <http://www.efsa.europa.eu/en/scdocs/scdoc/50e.htm>.
- Kagabu, S., Nishimura, K., Naruse, Y., & Ohno, I. (2008). Insecticidal and neuroblocking potencies of variants of the thiazolidine moiety of thiacloprid and quantitative relationship study for the key neonicotinoid pharmacophore. *Journal of Pesticide Science*, 33(1), 58–66.
- Kar, S., & Roy, K. (2012). QSAR of phytochemicals for the design of better drugs. *Expert Opinion on Drug Discovery*, 7(10), 877–902.
- Kar, S., Gajewicz, A., Puzyn, T., Roy, K., & Leszczynski, J. (2014). Periodic table-based descriptors to encode cytotoxicity profile of metal oxide nanoparticles: A mechanistic QSTR approach. *Ecotoxicology and Environmental Safety*, 107, 162–169.
- Kaur, I., & Geetha, T. (2006). Screening methods for antioxidants-A review. *Mini-Reviews in Medicinal Chemistry*, 6(3), 305–312.
- Kiwamoto, R., Spenkelink, A., Rietjens, I. M. C. M., & Punt, A. (2015). An integrated QSAR-PBK/D modelling approach for predicting detoxification and DNA adduct formation of 18 acyclic food-borne α , β -unsaturated aldehydes. *Toxicology and Applied Pharmacology*, 282 (1), 108–117.
- Kroes, R., Renwick, A. G., Cheeseman, M., Kleiner, J., Mangelsdorf, I., Piersma, A., ..., Würtzen, G. (2004). Structure-based thresholds of toxicological concern (TTC): Guidance for application to substances present at low levels in the diet. *Food and Chemical Toxicology*, 42(1), 65–83.
- Lamberth, C., Jeanmart, S., Luksch, T., & Plant, A. (2013). Current challenges and trends in the discovery of agrochemicals. *Science*, 341(6147), 742–746.

- Le Maux, S., Nongonierma, A. B., & FitzGerald, R. J. (2015). Improved short peptide identification using HILIC–MS/MS: Retention time prediction model based on the impact of amino acid position in the peptide sequence. *Food Chemistry*, *173*, 847–854.
- Lei, B., Li, J., Lu, J., Du, J., Liu, H., & Yao, X. (2009). Rational prediction of the Herbicidal activities of novel protoporphyrinogen oxidase inhibitors by quantitative structure-activity relationship model based on Docking-Guided active conformation. *Journal of Agricultural and Food Chemistry*, *57*(20), 9593–9598.
- Li, Y.-W., & Li, B. (2013). Characterization of structure–antioxidant activity relationship of peptides in free radical systems using QSAR models: Key sequence positions and their amino acid properties. *Journal of Theoretical Biology*, *318*, 29–43.
- Li, J., Ju, X. L., & Jiang, F. C. (2008). Pharmacophore model for neonicotinoid insecticides. *Chinese Chemical Letters*, *19*(5), 619–622.
- Li, Z. G., Chen, K. X., Xie, H. Y., & Gao, J. R. (2009a). Quantitative structure-property relationship studies on amino acid conjugates of Jasmonic acid as defense signaling molecules. *Journal of Integrative Plant Biology*, *51*(6), 581–592.
- Li, Z. G., Chen, K. X., Xie, H. Y., Chen, K. Y., & Shen, D. L. (2009b). QSAR analysis of jasmonates as novel plant growth regulators. *Chinese Journal of Pesticide Science*, *11*, 166–175.
- Li, Y.-W., Li, B., He, J., & Qian, P. (2011). Quantitative structure–activity relationship study of antioxidative peptide by using different sets of amino acids descriptors. *Journal of Molecular Structure*, *998*(1–3), 53–61.
- Li, D., Du, S., Tan, W., & Duan, H. (2015). Computational insight into the structure–activity relationship of novel n-substituted phthalimides with gibberellin-like activity. *Journal of Molecular Modeling*, *21*(10), 271.
- Liu, Y.-X., Wei, D.-G., Zhu, Y.-R., Liu, S.-H., Zhang, Y.-L., Zhao, Q.-Q., ..., Wang, Q.-M. (2008). Synthesis, Herbicidal activities, and 3D-QSAR of 2-Cyanoacrylates containing aromatic Methylamine Moieties. *Journal of Agricultural and Food Chemistry*, *56*(1), 204–212.
- Liu, G.-Y., Ju, X.-L., Cheng, J., & Liu, Z.-Q. (2010). 3D-QSAR studies of insecticidal anthranilic diamides as ryanodine receptor activators using CoMFA, CoMSIA and DISCOtech. *Chemosphere*, *78*(3), 300–306.
- Liu, X.-H., Xu, X.-Y., Tan, C.-X., Weng, J.-Q., Xin, J.-H., & Chen, J. (2015). Synthesis, crystal structure, herbicidal activities and 3D-QSAR study of some novel 1, 2,4-triazolo[4, 3- a] pyridine derivatives. *Pest Management Science*, *71*(2), 292–301.
- Loso, M. R., Benko, Z., Buysse, A., Johnson, T. C., Nugent, B. M., Rogers, R. B., ..., Zhu, Y. (2016). SAR studies directed toward the pyridine moiety of the sap-feeding insecticide sulfoxaflor (Isoclast™ active). *Bioorganic and Medicinal Chemistry*, *24*(3), 378–382.
- Lu, G.-N., Dang, Z., Tao, X.-Q., Chen, X.-P., Yi, X.-Y., & Yang, C. (2007). Quantitative structure-activity relationships for enzymatic activity of chloroperoxidase on metabolizing organophosphorus pesticides. *QSAR and Combinatorial Science*, *26*(2), 182–188.
- Martinez-Mayorga, K., & Medina-Franco, J. L. (2009). *Chemoinformatics-applications in food chemistry* (Vol. 58). Burlington: Elsevier.
- Mayer, J., Cheeseman, M. A., & Twaroski, M. L. (2008). Structure-activity relationship analysis tools: Validation and applicability in predicting carcinogens. *Regulatory Toxicology and Pharmacology*, *50*(1), 50–58.
- McCord, M. J. (2004). Therapeutic control of free radicals. *Drug Discovery Today*, *9*(18), 781–782.
- Mitra, I., Saha, A., & Roy, K. (2010a). Chemometric modeling of free radical scavenging activity of flavone derivatives. *European Journal of Medicinal Chemistry*, *45*(11), 5071–5079.
- Mitra, I., Saha, A., & Roy, K. (2010b). Pharmacophore mapping of arylamino-substituted benzo [b]thiophenes as free radical scavengers. *Journal of Molecular Modeling*, *16*(10), 1585–1596.
- Mitra, I., Saha, A., & Roy, K. (2011). Chemometric QSAR modeling and in silico design of antioxidant NO donor Phenols. *Scientia Pharmaceutica*, *79*(1), 31–57.

- Mitra, I., Saha, A., & Roy, K. (2012a). Development of multiple QSAR models for consensus predictions and unified mechanistic interpretations of the free-radical scavenging activities of chromone derivatives. *Journal of Molecular Modeling*, 18(5), 1819–1840.
- Mitra, I., Saha, A., & Roy, K. (2012b). In silico development, validation and comparison of predictive QSAR models for lipid peroxidation inhibitory activity of cinnamic acid and caffeic acid derivatives using multiple chemometric and cheminformatics tools. *Journal of Molecular Modeling*, 18(8), 3951–3967.
- Mitra, I., Saha, A., & Roy, K. (2013a). Predictive modeling of Antioxidant Coumarin derivatives using multiple approaches: Descriptor based QSAR, 3D-Pharmacophore mapping, and HQSAR. *Scientia Pharmaceutica*, 81(1), 57–80.
- Mitra, I., Saha, A., & Roy, K. (2013b). Quantification of contributions of different molecular fragments for antioxidant activity of coumarin derivatives based on QSAR analyses. *Canadian Journal of Chemistry*, 91(6), 428–441.
- Mitra, I., Saha, A., & Roy, K. (2013c). Predictive chemometric modeling of DPPH free radical-scavenging activity of azole derivatives using 2D- and 3D-quantitative structure–activity relationship tools. *Future Medicinal Chemistry*, 5(3), 261–280.
- Musialik, M., & Litwinienko, G. (2005). Scavenging of dpph* radicals by vitamin E is accelerated by its partial Ionization: The role of Sequential proton loss electron transfer. *Organic Letters*, 7(22), 4951–4954.
- Nakagawa, Y. (2007). Structure–activity relationship and mode of action study of insect growth regulators. *Journal of Pesticide Science*, 32(2), 135–136.
- Netzeva, T. I., Aptula, A. O., Benfenati, E., Cronin, M. T. D., Gini, G., Lessigiarska, I., ..., Schüürmann, G. (2005). Description of the electronic structure of organic chemicals using semiempirical and ab Initio methods for development of Toxicological QSARs. *Journal of Chemical Information and Modeling*, 45(1), 106–114.
- Niraj, R. R. K., Saini, V., & Kumar, A. (2015). QSAR analyses of organophosphates for insecticidal activity and its in-silico validation using molecular docking study. *Environmental Toxicology and Pharmacology*, 40(3), 886–894.
- Nishimura, K., Kiriya, K., & Kagabu, S. (2006). Quantitative structure–activity relationships of imidacloprid and its analogs with substituents at the C5 position on the pyridine ring in the neuroblocking activity. *Journal of Pesticide Science*, 31(2), 110–115.
- NRC. (2007). *Toxicity testing in the 21st Century; a vision and a strategy*. Washington, D.C.: National Academy Press.
- OECD. (2009). Guidance Document on the Definition of Residue. Series on Testing and Assessment No. 63 and Series on Pesticides No. 31. 28 July 2009. Retrieved August 18, 2016, from <http://www.oecd.org/>.
- OECD. (2014). Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, OECD Series on Testing and Assessment, No. 69, OECD Publishing, Paris. Retrieved August 18, 2016, from <http://dx.doi.org/10.1787/9789264085442-en>.
- Pérez-Garrido, A., Helguera, A. M., Morillas Ruiz, J. M., & Zafriilla Rentero, P. (2012). Topological sub-structural molecular design approach: Radical scavenging activity. *European Journal of Medicinal Chemistry*, 49, 86–94.
- Perkins, R., Fang, H., Tong, W., & Welsh, W. J. (2003). Quantitative structure-activity relationship methods: perspectives on drug discovery and toxicology. *Environmental Toxicology and Chemistry*, 22(8), 1666–1679.
- Prouillac, C., Vicendo, P., Garrigues, J.-C., Poteau, R., & Rima, G. (2009). Evaluation of new thiadiazoles and benzothiazoles as potential radioprotectors: Free radical scavenging activity in vitro and theoretical studies (QSAR, DFT). *Free Radical Biology and Medicine*, 46(8), 1139–1148.
- Rackova, L., Firakova, S., Kostalova, D., Stefek, M., Sturdik, E., & Majekova, M. (2005). Oxidation of liposomal membrane suppressed by flavonoids: Quantitative structure–activity relationship. *Bioorganic and Medicinal Chemistry*, 13(23), 6477–6484.

- Rastija, V., & Medić-Šarić, M. (2009). QSAR study of antioxidant activity of wine polyphenols. *European Journal of Medicinal Chemistry*, *44*(1), 400–408.
- Ray, S., De, K., Sengupta, C., & Roy, K. (2008). QSAR study of lipid peroxidation-inhibition potential of some phenolic antioxidants. *Indian Journal of Biochemistry and Biophysics*, *45*, 198–205.
- Reino, J. L., Saiz-Urra, L., Hernández-Galán, R., Arán, V. J., Hitchcock, P. B., Hanson, J. R., ... Collado, I. G. (2007). Quantitative structure–antifungal activity relationships of some benzohydrazides against *Botrytis cinerea*. *Journal of Agricultural and Food Chemistry*, *55* (13), 5171–5179.
- Rojas, C., Duchowicz, P. R., Tripaldi, P., & Diez, R. P. (2015). QSPR analysis for the retention index of flavors and fragrances on a OV-101 column. *Chemometrics and Intelligent Laboratory Systems*, *140*, 126–132.
- Rojas, C., Tripaldi, P., & Duchowicz, P. R. (2016). A new QSPR study on relative sweetness. *International Journal of Quantitative Structure-Property Relationships*, *1*(1), 78–93.
- Rouhollahi, A., Ghasemi, J., & Babae, E. (2010). Quantitative structure activity relationship modeling of environmentally important Diphenyl ether herbicides using MLR and PLS. *Current Analytical Chemistry*, *6*(1), 3–10.
- Roy, K., & Mitra, I. (2009). Advances in quantitative structure-activity relationship models of antioxidants. *Expert Opinion on Drug Discovery*, *4*(11), 1157–1175.
- Roy, K., & Paul, S. (2009). Exploring 2D and 3D QSARs of 2, 4-Diphenyl-1, 3-oxazolines for Ovicidal activity against *Tetranychus urticae*. *QSAR and Combinatorial Science*, *28*(4), 406–425.
- Roy, K., & Paul, S. (2010a). Docking and 3D-QSAR studies of acetohydroxy acid synthase inhibitor sulfonylurea derivatives. *Journal of Molecular Modeling*, *16*(5), 951–964.
- Roy, K., & Paul, S. (2010b). Docking and 3D QSAR studies of protoporphyrinogen oxidase inhibitor 3H-pyrazolo[3, 4-d][1, 2,3]triazin-4-one derivatives. *Journal of Molecular Modeling*, *16*(1), 137–153.
- Roy, K., Kar, S., & Das, R. N. (2015a). *A primer on QSAR/QSPR modeling: Fundamental concepts* (SpringerBriefs in Molecular Science), Springer.
- Roy, K., Kar, S., & Das, R. N. (2015b). *Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment*. Elsevier Science.
- Ruark, C. D., Hack, C. E., Robinson, P. J., Anderson, P. E., & Gearhart, J. M. (2013). Quantitative structure–activity relationships for organophosphates binding to acetylcholinesterase. *Archives of Toxicology*, *87*(2), 281–289.
- Russell, W. M. S., & Burch, R. L. (1959). *The principles of humane experimental technique*, London, UK.
- Saavedra, L. M., Ruiz, D., Romanelli, G. P., & Duchowicz, P. R. (2015). Quantitative structure-antifungal activity relationships for cinnamate derivatives. *Ecotoxicology and Environmental Safety*, *122*, 521–527.
- Samee, W., Nunthanavanit, P., & Ungwitayatorn, J. (2008). 3D-QSAR investigation of synthetic antioxidant chromone derivatives by molecular field analysis. *International Journal of Molecular Sciences*, *9*(3), 235–246.
- Samghani, K., & HosseinFatemmi, M. (2016). Developing a support vector machine based QSPR model for prediction of half-life of some herbicides. *Ecotoxicology and Environmental Safety*, *129*, 10–15.
- Seifert, J. (2016). The structural requirements of organophosphorus insecticides (OPI) for reducing chicken embryo NAD + content in OPI-induced teratogenesis in chickens. *Pesticide Biochemistry and Physiology*, *129*, 43–48.
- Slavov, S., Gini, G., & Benfenati, E. (2008). QSAR trout toxicity models on aromatic pesticides. *Journal of Environmental Science and Health, Part B*, *43*(8), 633–637.
- Smith, R. L., Cohen, S. M., Doull, J., Feron, V. J., Goodman, J. I., Marnett, L. J., ..., Adams, T. B. (2005). A procedure for the safety evaluation of natural flavor complexes used as ingredients in food: Essential oils. *Food and Chemical Toxicology*, *43*(3), 345–363.

- Song, J. S., Moon, T., Nam, K. D., Lee, J. K., Hahn, H.-G., Choi, E.-J., et al. (2008). Quantitative structural–activity relationship (QSAR) study for fungicidal activities of thiazoline derivatives against rice blast. *Bioorganic and Medicinal Chemistry Letters*, 18(6), 2133–2142.
- Speck-Planche, A., Kleandrova, V. V., & Scotti, M. T. (2012). Fragment-based approach for the in silico discovery of multi-target insecticides. *Chemometrics and Intelligent Laboratory Systems*, 111(1), 39–45.
- Sprous, D. G., & Salemme, F. R. (2007). A comparison of the chemical properties of drugs and FEMA/FDA notified GRAS chemical compounds used in the food industry. *Food and Chemical Toxicology*, 45(8), 1419–1427.
- Takaki, K., Wade, A. J., & Collins, C. D. (2015). Assessment and improvement of biotransfer models to cow's milk and beef used in exposure assessment tools for organic pollutants. *Chemosphere*, 138, 390–397.
- Tan, J., Tian, F., Lv, Y., Liu, W., Zhong, L., Liu, Y., et al. (2013). Integration of QSAR modelling and QM/MM analysis to investigate functional food peptides with antihypertensive activity. *Molecular Simulation*, 39(12), 1000–1006.
- Tarko, L., Lupescu, I., & Constantinescu-Groprosil, D. (2006). QSAR studies on amino-succinamic acid derivatives sweeteners. *Arkivoc*, 13, 22–40.
- Tikhonova, I. G., Baskin, I. I., Palyulin, V. A., & Zefirov, N. S. (2004). Virtual screening of organic molecule databases. Design of focused libraries of potential ligands of NMDA and AMPA receptors. *Russian Chemical Bulletin*, 53(6), 1335–1344.
- Tilaoui, L., Schilter, B., Tran, L.-A., Mazzatorra, P., & Grigorov, M. (2007). Integrated computational methods for prediction of the lowest observable adverse effect level of food-borne molecules. *QSAR and Combinatorial Science*, 26(1), 102–108.
- Tromelin, A., & Guichard, E. (2003). Use of catalyst in a 3D-QSAR study of the interactions between flavor compounds and β -lactoglobulin. *Journal of Agricultural and Food Chemistry*, 51(7), 1977–1983.
- Vajragupta, O., Boonchoong, P., & Wongkrajang, Y. (2000). Comparative quantitative structure–activity study of radical scavengers. *Bioorganic and Medicinal Chemistry*, 8(11), 2617–2628.
- Valko, M., Jomova, K., Rhodes, C. J., Kuča, K., & Musílek, K. (2016). Redox- and non-redox-metal-induced formation of free radicals and their role in human disease. *Archives of Toxicology*, 90, 1–37.
- Velkov, Z. A., Kolev, M. K., & Tadjer, A. V. (2007). Modeling and statistical analysis of DPPH scavenging activity of Phenolics. *Collection of Czechoslovak Chemical Communications*, 72 (11), 1461–1471.
- Vencill, W. K. (2002). *Herbicide handbook* (8th ed., p. 493). Lawrence KS: Weed Science Society of America.
- Vepuri, S. B., Tawari, N. R., & Degani, M. S. (2007). Quantitative structure-activity relationship study of some aspartic acid analogues to correlate and predict their sweetness potency. *QSAR and Combinatorial Science*, 26(2), 204–214.
- Vidhyasekaran, P. (2004). Chemical control-virus diseases. *Concise Encyclopedia of Plant Pathology, Food Products Press® and The Haworth Reference Press* (pp. 413–416). NY: Binghamton.
- Vinholes, J., Rudnitskaya, A., Gonçalves, P., Martel, F., Coimbra, M. A., & Rocha, S. M. (2014). Hepatoprotection of sesquiterpenoids: A quantitative structure–activity relationship (QSAR) approach. *Food Chemistry*, 146, 78–84.
- Wang, J.-G., Li, Z.-M., Ma, N., Wang, B.-L., Jiang, L., Pang, S. S., ..., Duggleby, R. G. (2005). Structure-activity relationships for a new family of sulfonylurea herbicides. *Journal of Computer-Aided Molecular Design*, 19(11), 801–820.
- Wang, G., Li, Y., Liu, X., & Wang, Y. (2009). Understanding the aquatic toxicity of pesticide: Structure-activity relationship and molecular descriptors to distinguish the ratings of toxicity. *QSAR and Combinatorial Science*, 28(11–12), 1418–1431.
- Wang, J., Tan, H., Li, Y., Ma, Y., Li, Z., & Guddat, L. W. (2011). Chemical synthesis, in vitro Acetohydroxyacid Synthase (AHAS) inhibition, Herbicidal activity, and computational studies of Isatin derivatives. *Journal of Agricultural and Food Chemistry*, 59(18), 9892–9900.

- Wang, Y., Shao, Y., Fan, L., Yu, X., Zhi, X., Yang, C., ..., Xu, H. (2012). Synthesis and quantitative structure–activity relationship (QSAR) study of novel isoxazoline and oxime derivatives of podophyllotoxin as insecticidal agents. *Journal of Agricultural and Food Chemistry*, 60(34), 8435–8443.
- Wang, J.-H., Liu, Y.-L., Ning, J.-H., Yu, J., Li, X.-H., & Wang, F.-X. (2013). Is the structural diversity of tripeptides sufficient for developing functional food additives with satisfactory multiple bioactivities? *Journal of Molecular Structure*, 1040, 164–170.
- Wang, M.-J., Zhao, X.-B., Wu, D., Liu, Y.-Q., Zhang, Y., Nan, X., ..., Yan, L.-T. (2014). Design, synthesis, crystal structure, insecticidal activity, molecular docking, and QSAR studies of novel N 3 -substituted imidacloprid derivatives. *Journal of Agricultural and Food Chemistry*, 62(24), 5429–5442.
- Waxman, M. F. (1998). *The agrochemical and pesticides safety handbook*. CRC Press.
- Wei, S., Ji, Z., Zhang, H., Zhang, J., Wang, Y., & Wu, W. (2011). Isolation, biological evaluation and 3D-QSAR studies of insecticidal/narcotic sesquiterpene polyol esters. *Journal of Molecular Modeling*, 17(4), 681–693.
- Wright, J. S., Johnson, E. R., & DiLabio, G. A. (2001). Predicting the activity of phenolic antioxidants: Theoretical method, analysis of substituent effects, and application to major families of antioxidants. *Journal of the American Chemical Society*, 123(6), 1173–1183.
- Wu, J., & Aluko, R. E. (2007). Quantitative structure-activity relationship study of bitter di- and tri-peptides including relationship with angiotensin i-converting enzyme inhibitory activity. *Journal of Peptide Science*, 13(1), 63–69.
- Xia, S., Feng, Y., Cheng, J.-G., Luo, H.-B., & Li, Z. (2014). QAAR exploration on pesticides with high solubility: An investigation on sulfonylurea herbicide dimers formed through π - π stacking interactions. *Chinese Chemical Letters*, 25(7), 973–977.
- Xu, H., Wang, J., Sun, H., Lv, M., Tian, X., Yao, X., et al. (2009). Semisynthesis and quantitative structure-activity relationship (QSAR) study of novel aromatic esters of 4'-Demethyl-4-deoxypodophyllotoxin as insecticidal agents. *Journal of Agricultural and Food Chemistry*, 57(17), 7919–7923.
- Yang, X., Chong, Y., Yan, A., & Chen, J. (2011). In-silico prediction of sweetness of sugars and sweeteners. *Food Chemistry*, 128(3), 653–658.
- Yang, C., Shao, Y., Zhi, X., Huan, Q., Yu, X., Yao, X., et al. (2013). Semisynthesis and quantitative structure–activity relationship (QSAR) study of some cholesterol-based hydrazone derivatives as insecticidal agents. *Bioorganic and Medicinal Chemistry Letters*, 23(17), 4806–4812.
- Yuan, M., Liu, B., Liu, E., Sheng, W., Zhang, Y., Crossan, A., ..., Wang, S. (2011). Immunoassay for Phenylurea herbicides: Application of molecular modeling and quantitative structure–activity relationship analysis on an antigen–antibody interaction study. *Analytical Chemistry*, 83(12), 4767–4774.
- Zhao, W.-G., Wang, J.-G., Li, Z.-M., & Yang, Z. (2006). Synthesis and antiviral activity against tobacco mosaic virus and 3D-QSAR of α -substituted-1, 2,3-thiadiazoleacetamides. *Bioorganic and Medicinal Chemistry Letters*, 16(23), 6107–6111.
- Zhong, M., Chong, Y., Nie, X., Yan, A., & Yuan, Q. (2013). Prediction of sweetness by multilinear regression analysis and support vector machine. *Journal of Food Science*, 78(9), S1445–S1450.
- Zhou, P., Yang, C., Ren, Y., Wang, C., & Tian, F. (2013). What are the ideal properties for functional food peptides with antihypertensive effect? A computational peptidology approach. *Food Chemistry*, 141(3), 2967–2973.
- Zhu, Y., Wu, C., Li, H., Zou, X., Si, X., Hu, F., et al. (2007). Design, synthesis, and quantitative structure-activity relationship study of herbicidal analogues of Pyrazolo[5, 1- d][1, 2, 3, 5] tetrazin-4(3 H)ones. *Journal of Agricultural and Food Chemistry*, 55(4), 1364–1369.
- Zimdahl, R. L. (2007). *Properties and uses of herbicides. Fundamentals of weed science* (pp. 395–435). London: Elsevier.

Quantitative Structure-Epigenetic Activity Relationships

Mario Omar García-Sánchez, Maykel Cruz-Monteagudo
and José L. Medina-Franco

Abstract The relevance of epigenetic drug discovery has increased during the past few years as revealed by the augmenting number of related publications and the amount of structure-epigenetic activity data in compound databases. This chapter discusses the current status of epigenetic target-based therapies. It is also analyzed the progress of quantitative structure-activity relationship (QSAR) models developed for compound databases screened with epigenetic targets. A special emphasis is made on compounds directed to inhibitors of DNA methyltransferases, one of the first epigenetic target families associated with therapeutic potential. Novel approaches applied to develop models for inhibitors of bromodomains, other epigenetic target families with high relevance in modern drug discovery programs, are also discussed. The chapter analyses epigenetic activity landscape modeling, activity cliffs, and activity cliff generators and their relevance to develop QSAR models. Computational methods applied to elucidate Quantitative Structure-Epigenetic Activity Relationships are in line with the increasing and developing research area of Epi-informatics.

Keywords Activity cliffs · Activity landscape · Bromodomains · DNA methyltransferase · Epigenetics · Epi-informatics · HDAC · SEARS

M.O. García-Sánchez · J.L. Medina-Franco (✉)
Facultad de Química, Departamento de Farmacia, Universidad Nacional
Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico
e-mail: medinajl@unam.mx; jose.medina.franco@gmail.com

M.O. García-Sánchez
e-mail: llueve_mario@hotmail.com

M. Cruz-Monteagudo
Center for Computational Science, University of Miami, 33136 Miami, FL, USA
e-mail: gmailkelcm@yahoo.es

M. Cruz-Monteagudo
Faculdade de Ciências, CIQUP/Departamento de Química e Bioquímica,
Universidade do Porto, Porto 4169-007, Portugal

M. Cruz-Monteagudo
Facultad de Ciências, REQUIMTE, Departamento de Química e Bioquímica,
Universidade do Porto, Porto 4169-007, Portugal

List of abbreviations

ACG	Activity Cliffs Generators
ALM	Activity Landscape Modeling
BRD	Bromodomain
DNMT	DNA Methyltransferase
ERCS	Epigenetic Relevant Chemical Space
FDA	Food and Drug Administration
HDAC	Histone lysine Deacetylase
ISMs	Instances that should be Misclassified
MMP	Matched Molecular Pairs
MODI	Modelability Index
NSG	Network-like Similarity Graphs
PLIF	Protein-Ligand Interaction Fingerprint
PLM	Property Landscape Modeling
QSAR	Quantitative Structure-Activity Relationship
SALI	Structure Activity Landscape Index
SAR	Structure-Activity Relationship
SAS	Structure-Activity Similarity
SEARSSEARS	Structure-Epigenetic Activity Relationships
SARI	Structure-Activity Relationship Index
SmAR	Structure-multiple Activity Relationship
SVMs	Support Vector Machines

1 Epigenetics and Computational Approaches

1.1 Relevance of Epigenetics and Major Epigenetic Targets

Epigenetics is the study of the elements that participate in the regulation of the nucleosome-chromatin as determinants of gene expression. The term “Epigenetics” itself was created by Conrad Waddington while he was attempting to describe “the interactions of genes with their environment, which brings the phenotype into being (Waddington 2012). Since the early 1940’s the definition of epigenetics has been changing to include the analysis of heritable phenotypic traits that result from modifications to a chromosome that do not modify the basic genetic code (Berger et al. 2009).

Major epigenetic targets of therapeutic interest are histone lysine deacetylases (HDACs), bromodomains (BRDs), and DNA methyltransferases (DNMTs). In depth reviews of these and other epigenetic targets have been published elsewhere (Dueñas-González et al. 2016). Currently six epigenetic drugs have been approved by the Food and Drug Administration (FDA) of the United States for cancer indications: two DNA demethylating and four histone deacetylase inhibitors

(Dueñas-González et al. 2016). Figure 1 summarizes the agents and year of approval of these epigenetic drugs. The first four FDA-approved drugs were identified based on observations of cell phenotypes; several years later their biochemical targets discovered. It is worth noting that belinostat and panobinostat were the first drugs developed after their molecular targets were uncovered.

Epigenetic drug discovery is rapidly growing in both industry and academia. Indeed, the increasing interest in epigenetic drug and probe discovery is reflected in the growing number of publications. For example, Arguelles et al. have recently showed the increased number of epigenetic-related publications in the period 2010–2015 (Arguelles et al. 2016). Similarly, the structure-epigenetic activity information stored in public molecular databases has increased significantly. Moreover, epigenetic targets are now abundant in drug discovery pipelines of multiple biotechnology companies. In order to accelerate the identification and development of novel epigenetic drugs and probes, several technological advances have been made to screen compound databases using low and high throughput assays (Zheng 2015).

The fact that the six drugs (Fig. 1) are approved against cancer by no means implies that epigenetic alterations are limited to malignant conditions. The study of epigenetics in many chronic diseases has rapidly evolved and now a large number of studies have been published on the epigenetic pathogenesis of these diseases. Table 1 summarizes representative diseases that may be approached with compounds directed to epigenetic targets (Dueñas-González et al. 2016; Alam et al. 2016).

The significance of epigenetic drug and probe development combined with the large amount of SEA requires computational approaches to organize and mine the experimental data to further advance the development of epi-drugs and epi-probes (Dueñas-González et al. 2016). One of the first steps to explore Structure-Epigenetic

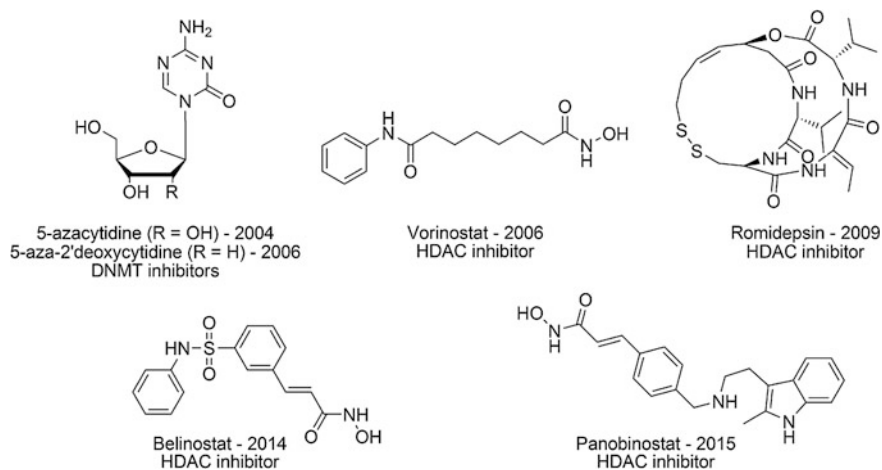


Fig. 1 Currently epigenetic drugs approved by the food and drug administration (FDA) of the United States

Table 1 Examples of diseases that may be approached with epigenetic therapies

General type of condition or disorder	Diseases
Neurodegenerative	Alzheimer's, parkinson's, huntington's, multiple sclerosis, epilepsy, schizophrenia, bipolar disorders and other psychiatric conditions
Respiratory	Asthma, chronic obstructive pulmonary disease, pulmonary hypertension, lung fibrosis
Cardiovascular	Atherosclerosis, coronary artery disease, heart failure
Autoimmune	Systemic lupus erythematosus, rheumatoid arthritis, systemic sclerosis and Sjögren's syndrome

Activity RelationshipS (SEARS) is to characterize the chemical space of epigenetic compounds, i.e., to quantify the chemical diversity, scaffold content, and their coverage in chemical space. The first attempts to characterize the so-called Epigenetic Relevant Chemical Space (ERCS) have been published. For instance, Fernández de-Gortari and Medina-Franco (2015) assembled and curated a molecular database of inhibitors of DNMT subtype1 (DNMT1). Compounds were analyzed in terms of physicochemical properties, structural diversity, coverage of chemical space, and scaffold diversity. In a follow-up work, Prieto-Martinez et al. (2016) reported the first comprehensive chemoinformatic analysis of BRD and HDAC inhibitors available in the public domain. Similar to the work of Fernández de-Gortari, compounds were analyzed in terms of physicochemical properties, molecular fingerprints and structural scaffolds. In addition, molecular complexity was measured for inhibitors of DNMTs, HDAC and DNMT1s. It was concluded that, despite the fact that several compounds have been screened against these molecular targets, there are not yet molecular scaffold with large enrichment factors, e.g., epigenetic privileged scaffolds remain to be identified. It was also concluded that there is a need to increase the molecular complexity of screening libraries to explore novel regions in chemical space for epigenetic targets. These two works were aimed to explore parts of the ERCS in a systematic manner (Fernandez-de Gortari and Medina-Franco 2015; Prieto-Martinez et al. 2016).

1.2 Overview of Computational Studies for Epigenetic Drug Development

Computational structure—and ligand-based approaches have had numerous applications to identify and develop drugs and probes for epigenetic targets. Likewise, chemoinformatic methods are playing a key role to organize, mine, and visualize the information generated in epigenetic projects (Medina-Franco and Yoo 2016). For instance, in silico methods have been used to analyze and predict ligand-target interactions and describe in a systematic manner SARs (Structure Activity

Relationships)/SEARS (see below). As discussed in this chapter, the specific methods applied for a particular problem depend on the experimental information available as well as the goals of the study (Méndez-Lucio 2016). Overall, major computational approaches used for epigenetic drug and probe development include, but are not limited to, virtual screening, molecular dynamics, docking, pharmacophore modeling and quantitative structure-activity relationships (QSAR).

Ideally, QSAR models could be interpreted at the molecular level through the rationalization of protein–ligand interactions, i.e., establishing structure–protein–ligand interaction relationships. Certainly, it has been pointed out that understanding protein–ligand interactions is at the core of molecular recognition and they are implicated with several practical applications including structure-based design, docking, virtual screening, and clustering of protein-ligand complexes (Medina-Franco et al. 2014). As discussed below, protein-ligand interactions can be studied using protein-ligand interaction fingerprints (PLIFs) also termed structural interaction fingerprints. PLIFs are designed to ‘capture a 1D representation of the interactions between ligand and protein either in complexes of known structure or in docked poses’ (Brewerton 2008; Desaphy et al. 2013). In epigenetics and other research areas, PLIFs are quite useful to analyze large amounts of structural data and get insights into different areas, for instance (Desaphy et al. 2013; Méndez-Lucio 2016):

- If similar binding sites recognize similar ligands.
- If protein-ligand interaction patterns are conserved across target families.
- If different ligand structures or substructures have similar interaction patterns with a single target.

PLIFs and their applications to identify hot spots have been reviewed recently (Medina-Franco et al. 2014).

2 Overview of QSAR

The main idea of any QSAR model is to establish quantitative relationships between a biological endpoint with the chemical structures that are encoded in a quantitative manner using molecular descriptors. Overall, the practical application of such models is two-fold:

- Explain in a retrospective manner the biological endpoint of a series of compounds, and
- Predict the biological endpoint of a novel series of molecules.

Since the first QSAR model published by Hansch et al. in the early 1960’s, QSAR models have experienced significant modifications (Kubinyi 2002). In the early models, biological activity was correlated with simple structural parameters as molecular descriptors such as logP and Hammett’s σ , using linear regression. Today, more complex descriptors are employed which are applied to larger data sets

using not only linear regression but sophisticated machine learning techniques (Méndez-Lucio 2016). With the increasing importance of polypharmacology, screening data sets are tested across more than one molecular target. Therefore, QSAR models are being adapted to model quantitatively structure-multiple activity relationships (SmAR) (Waddell and Medina-Franco 2012; Medina-Franco and Waddell 2012).

3 Integration of Activity Landscape Modeling (ALM) and QSAR

Data sets of molecules screened across epigenetic targets represent reach starting points to analyze SARs in a descriptive and/or predictive manner. Examples of comprehensive molecular databases that store SEA are ChEMBL (Papadatos and Overington 2014), Binding Database (Liu et al. 2007), HEMD (Huang et al. 2012a), and Chromohub (Liu et al. 2012). Table 2 summarizes further information of each database.

It has been recognized that before developing QSAR models and making predictions, it is convenient to first understand the SAR. In particular, understanding the SAR of large data sets can be done through the systematic application of descriptive methods such as activity landscape modeling (ALM) (Medina-Franco et al. 2015b).

Table 2 Examples of molecular databases that store structure epigenetic-activity data

Database	Overall description	URL Link
HEMD: Human epigenetic enzyme and modulator database	'Central resource for the display, search, and analysis of the structure, function, and related annotation for human epigenetic enzymes and chemical modulators focused on epigenetic therapeutics'	http://mdl.shsmu.edu.cn/HEMD/
ChromoHub	'Online resource where users can map on phylogenetic trees disease associations, protein structures, chemical inhibitors, histone substrates, chromosomal aberrations and other types of data extracted from public repositories and the published literature'	http://www.thesgc.org/chromohub/
ChEMBL	'Manually curated chemical database of bioactive molecules with drug-like properties'	https://www.ebi.ac.uk/chembl/
Binding database	'Public, web-accessible database of measured binding affinities, focusing chiefly on the interactions of protein considered to be drug-targets with small, drug-like molecules'	https://www.bindingdb.org/

3.1 Activity Landscape Modeling and Activity Cliffs

The concept of activity landscapes was first introduced in the seminal work of Maggiora (Maggiora 2006). In that work Maggiora first defined a typical n -dimensional activity landscape as composed of an $(n-1)$ -dimensional chemical space where each dimension is described by a coordinate, which is generally defined by a single molecular descriptor or combination of descriptors. The n th dimension is defined by the activity space that is derived from the measured activity of each of the assayed compounds. Therefore, in three dimensions activity landscapes resemble the landscapes seen in nature (Fig. 2) and as so, represent a useful tool for exploring SARs (Maggiora 2006; Peltason and Bajorath 2007b; Bajorath et al. 2009; Peltason et al. 2010; Perez-Villanueva et al. 2010; Vogt et al. 2010; Iyer and Bajorath 2011; Bajorath 2012; Medina-Franco 2012; Mendez-Lucio et al. 2012b; Waddell and Medina-Franco 2012; Aguayo-Ortiz et al. 2014; Medina-Franco et al. 2015a; Méndez-Lucio et al. 2015).

In activity landscapes, smooth regions represent areas where gradual changes in chemical structure induce moderate changes in biological activity and are associated with continuous SARs. In contrast, rugged regions are associated with discontinuous SARs where small chemical modifications drastically change the biological response (Bajorath et al. 2009).

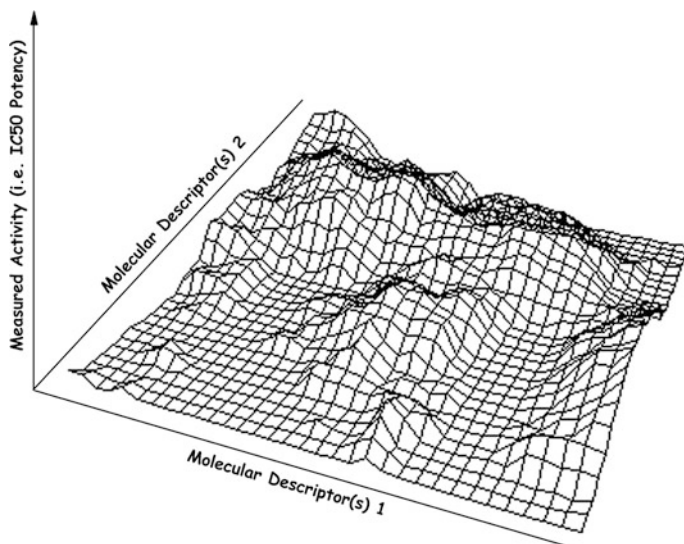


Fig. 2 Hypothetical three dimensional activity landscape. The plane defined by descriptors 1 and 2 is a low dimension—representation—of the chemical space. The relationship between the measured activity on the Z-axis and the chemical space gives rise to a structure-activity surface or activity landscape

While there are many areas where the activity landscape concept can provide guidance such as the pharmaceutical, chemical, food, agricultural and environmental applications (Wassermann and Bajorath 2010; Guha 2012b; Medina-Franco et al. 2015a), one of the main focuses of this concept relies on activity cliffs, irrespective of the task. Activity cliffs represent extreme forms of SAR discontinuity in activity landscapes and are formed by pairs of compounds with high structural similarity but unexpectedly high activity (or property) difference (Peltason and Bajorath 2011; Medina-Franco 2012; Stumpfe and Bajorath 2012a).

As deduced from its outlier nature, the activity cliffs influence on the activity landscape and the corresponding impact over activity landscapes based tasks can be significant, in a negative or positive way, depending on the task (Cruz-Monteagudo et al. 2014). For example, in drug discovery tasks, discontinuous SARs and activity cliffs provide the basis for lead optimization (Stumpfe and Bajorath 2012a; Dimova et al. 2013) while continuous SARs provides the basis for SAR progression analysis for medicinal chemists (Iyer et al. 2011). On the other hand, continuous and smooth SAR regions are prerequisites for the successful application of similarity-based methods for scaffold hopping or simply as predictive tools (Guha and Van Drie 2008b; Bajorath et al. 2009). These quantitative approaches rely on the Similarity Property Principle (Johnson and Maggiora 1990) that states that similar molecules should have similar activity, thus assuming the presence of continuous SARs. In contrast, in rugged and discontinuous SAR regions, the application of similarity-based methods is meaningless. Consequently, multiple definitions of activity cliffs have been introduced, adapting the essence of the phenomenon to different formalisms, depending of the task.

For instance, Guha introduced the Structure Activity Landscape Index (SALI) (Guha Van Drie 2008b), which is designed to identify activity cliffs and compounds representing key inflection points on activity landscapes. In a set of compounds, SALI assigns each compound pair a score that combines their pairwise similarity and the difference between their potency. The SALI approach is appropriate to detect activity cliffs in a dataset. However, the magnitude of the activity cliffs is not determined by the SALI metric since its values are compared on a relative scale. In a different approach, Stumpfe and Bajorath use discrete criteria to define activity cliffs including the applied similarity criterion, the potency measure and the magnitude of the potency difference. Stumpfe and Bajorath recommend considering a pair of compounds as an activity cliff only if a pre-established similarity criterion is satisfied; one compound in the pair has potency in the nanomolar range, and there is at least a 100-fold difference in potency between the two compounds (Stumpfe et al. 2013).

However, activity cliffs defined using similarity approaches are sometimes questioned by medicinal chemists because of their limited chemical interpretability (Wassermann and Bajorath 2010; Stumpfe and Bajorath 2012a). The application of the matched molecular pairs (MMP) formalism (Kenny and Sadowski 2004) addresses this issue (Hu et al. 2012a, b). A MMP is defined as a pair of compounds that only differ at a single site (represented by a substructure) such as a ring or an R-group. Thus, to classify a molecule pair as an MMP-cliff, the potency difference

required remains essentially the same as that applied in similarity-based definitions. Instead, the difference in size of the exchanged fragments and their size are restricted to a predefined maximum number of non-hydrogen atoms that guarantee the level of structural similarity expected for an activity cliff (Hu et al. 2012a, b).

Finally, activity cliffs have also been defined on the basis of consistently defined scaffolds and the presence of different scaffold/R-group relationships (Hu et al. 2012a, b; Aguayo-Ortiz et al. 2014; Medina-Franco et al. 2013) or by calculating the three dimensional similarity between compound binding modes observed in the X-ray structure of ligand/target complexes (Hu and Bajorath 2012; Hu et al. 2012a, b). An extensive review of the existing activity cliffs definitions is published elsewhere (Hu et al. 2013; Stumpfe et al. 2013).

3.2 Approaches for Activity Landscape Modeling

There are several approaches for ALM which can be classified according to the quantification of the SAR data, representation and method of visualization used (Bajorath et al. 2009; Bajorath 2012).

3.2.1 Structure-Activity Similarity (SAS) Maps

Structure-activity similarity (SAS) maps were one of the first approaches for ALM (Shanmugasundaram and Maggiora 2001) currently and actively used for SAR mapping and characterization (Mendez-Lucio et al. 2012a, b, 2015; Waddell and Medina-Franco 2012; Aguayo-Ortiz et al. 2014; Martinez-Mayorga et al. 2013, 2014; Yongye and Medina-Franco 2013; Navarrete-Vázquez and Méndez-Lucio 2015a). SAS maps are represented as two dimensional graphs that plot the similarity relationships between the chemical structure (usually plotted on the X-axis) and biological activity (usually represented by potency differences and plotted in the Y-axis) for each pair of molecules in a given dataset.

Basically, any structural representation and similarity metric can be used to compute structure similarity. On the other hand, the activity similarity can be computed in a relative scale (range scaled between 0 and 1) by considering the absolute activity difference of a pair of molecules in the range of activity values determined by the full set of molecules under study:

$$S_A(i, j) = 1 - \frac{|A_i - A_j|}{A_{max} - A_{min}} \quad (1)$$

where A_i and A_j are the activities of molecules i and j while A_{max} and A_{min} are the maximum and minimum activities, respectively.

A hypothetical SAS map is shown in Fig. 3a. As illustrated in the figure, these maps can be roughly divided into four quadrants in order to focus on particular

tasks according to the major SAR trend dominating the respective quadrant. The space determined by quadrant I, for instance, is particularly suitable for scaffold hopping and virtual screening since it is populated by pairs of molecules with low structure similarity and low potency difference (high activity similarity). On the contrary, the quadrant III is populated by pair of molecules with high structure similarity but large potency difference. Hence, this zone of the SAS map is appropriate for identifying activity cliffs, providing useful information for lead optimization tasks. Quadrant IV contains pairs of molecules with high structure similarity and low potency difference; thus, it is associated with a smooth or continuous SAR providing the basis for the meaningful application of similarity based predictive approaches such as QSAR modeling. Finally, quadrant II is in principle an uninformative region since it is dominated by pairs of molecules with low structure and activity similarities (Fig. 3a).

3.2.2 Structure Activity Landscape Index (SALI)

SALI (Guha and Van Drie 2008b) is a metric specifically developed to identify and quantify activity cliffs. SALI assigns each molecule pair a score that combines their pairwise similarity and the difference between their potency according to the expression:

$$SALI_{i,j} = \frac{|A_i - A_j|}{1 - sim(i,j)} \quad (2)$$

where $sim(i, j)$ is the similarity coefficient between the two molecules.

Just like SAS maps, any structural representation and similarity metric can be used to compute SALI. Therefore, pairs of molecules with high SALI values represent key inflection points on activity landscapes and consequently are considered as potential activity cliffs. The visualization of SALI values and its efficient use as an ALM approach can be conducted through a matrix representation termed as SALI heatmaps and the corresponding network representation derived from such matrices termed as SALI networks (Guha 2012a).

The simplest form to visualize the SALI values corresponding to a given set of molecules is to directly plot the corresponding SALI matrix as a heatmap, as shown in Fig. 3b. In SALI heatmaps X- and Y axes are ordered so that less active molecules are located towards the origin. SALI values in the heatmap are color coded (i.e., in a black to white scale) where light/dark blocks represent large/smaller SALI values pointing in this manner to potential activity cliffs.

Alternatively, the SALI matrix can be visualized through a network representation termed as SALI networks, as shown in Fig. 3b. For this type of visualization, an arbitrary SALI value must be previously specified as a threshold to select those pairs whose SALI value is greater. The common practice is to select a percentage of the maximum SALI value for the dataset that guarantees the significance of the cliff

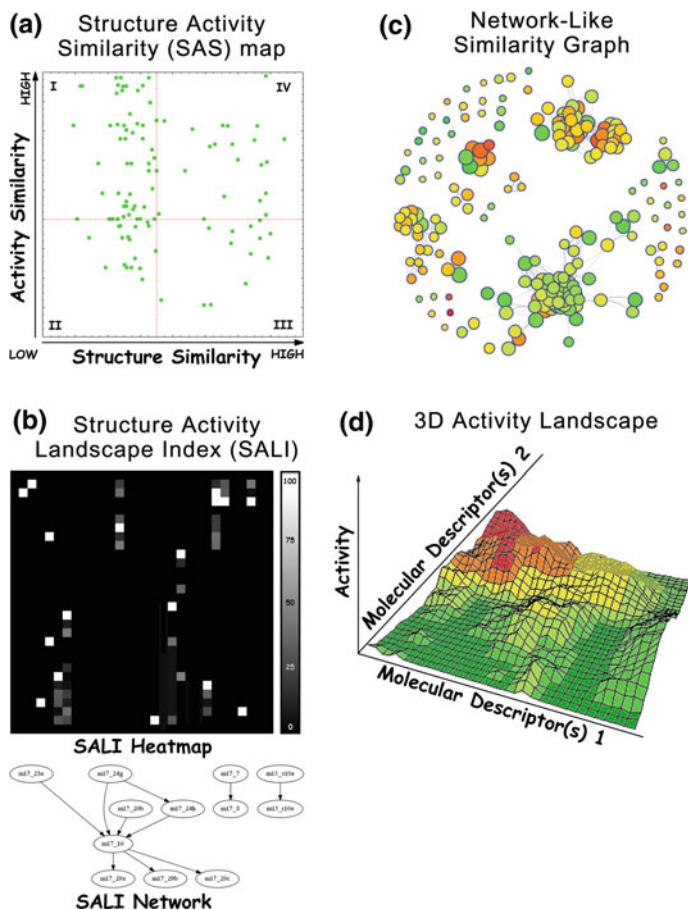


Fig. 3 Graphic representation of main activity landscape modeling approaches

pair identified and/or fit to the goals of the ALM analysis. The higher the SALI threshold, the higher is the significance of the activity cliffs identified; this gives rise to a sparser network which is easier to explore. On the contrary, at lower SALI thresholds, the activity cliffs identified are less significant, the resultant network is more densely populated and consequently, it is more complex its exploration.

3.2.3 Network-like Similarity Graphs (NSG)

The Network-like Similarity Graphs (NSG) (Wawer et al. 2008) approach for ALM relies on several global and local metrics for the quantification of the SAR encoded in a given dataset. The main metric behind NSGs is the structure-activity relationship index (SARI) (Peltason and Bajorath 2007b) which combines continuity

and discontinuity scores encoding global SAR continuity and discontinuity, respectively. The continuity score encodes the potency-weighted mean of pairwise compound dissimilarity, highlighting the presence of structurally diverse and potent compounds. The continuity score computed from raw data ($cont_{raw}$) is defined as:

$$cont_{raw} = \frac{\sum_{(i \neq j)} (weight(i,j) / (1 + sim(i,j)))}{\sum_{i \neq j} weight(i,j)} \quad (3)$$

$$weight(i,j) = \frac{A_i A_j}{1 + |A_i - A_j|} \quad (4)$$

On the other hand, the discontinuity score encodes the similarity-scaled average potency difference among ligand pairs that exceed a predefined similarity threshold and have a significant potency difference, highlighting the presence of activity cliffs. The discontinuity score computed from raw data ($disc_{raw}$) is defined as:

$$disc_{raw} = \text{mean}_{\{(i,j) | sim(i,j) > threshold_{sim}, |A_i - A_j| > 1\}} (|A_i - A_j| sim(i,j)) \quad (5)$$

Both, the continuity and discontinuity raw scores are normalized (range scaled between 0 and 1) and combined into the final SARI, as expressed below:

$$SARI = \frac{1}{2} (cont_{norm} + (1 - disc_{norm})) \quad (6)$$

Finally, SARI can be used to categorize the overall SAR present in a given dataset as continuous, discontinuous or heterogeneous according the high, low or medium SARI values obtained, respectively.

To reflect the potential participation of each compound in local activity cliffs and hence individual contributions to local and global SAR characteristics, a local variant of the discontinuity score can be applied to a given compound and its nearest neighbors:

$$disc(i)_{raw} = \text{mean}_{\{j \neq i | sim(i,j) > threshold_{sim}\}} (|A_i - A_j| sim(i,j)) \quad (7)$$

In NSGs, like with SAS maps and SALI applications, any structural representation and similarity metric can be used to compute the structure similarity of a given pair of compounds $sim(i, j)$. The structural similarity threshold ($threshold_{sim}$) is predefined according to the structural representation used (i.e. MACSS keys (Durant et al. 2002) or extended connectivity fingerprints (ECFP) (Rogers and Hahn 2010)). The biological activity is usually the corresponding potency measured as IC_{50} or K_i in nanomolar units (nM) and expressed in a log scale. So, a potency difference ($|A_i - A_j| \geq 1$) between a pair of compounds reflects a difference of one order of magnitude or higher and is usually considered as significant. Therefore, NSGs represent a compound dataset by showing all molecules and their similarity

relationships. As shown in Fig. 3c, NSGs are formal graphs in which nodes correspond to molecules. Pair-wise similarity relationships are represented by edges that connect individual nodes. Only molecule pairs that exceed a predefined similarity threshold are connected by an edge. To visualize the activity (potency) distribution, nodes are color coded by potency, applying a continuous spectrum from green (lowest) to red (highest potency). As commented above, the compound discontinuity score reflects SAR characteristics of individual molecules and is represented by node scaling reflecting the potency deviation of a compound from its structural neighbors. Large nodes represent compounds inducing a high discontinuity and vice versa (Wawer and Bajorath 2010, Stumpfe and Bajorath 2012b). Thus, it identifies molecules that introduce SAR discontinuity and activity cliffs. In NSGs, combinations of large red and green nodes connected by an edge are activity cliff markers that can be easily identified. NSGs are implemented in SARANEA (Lounkine et al. 2009), a freely available program that implements a graphical user interface to NSGs and NSG-based data mining techniques.

3.2.4 Three-Dimensional Activity Landscapes

The three dimensional activity landscape is the more intuitive but elaborated representation used in ALM. This type of activity landscape representation is obtained by dimension reduction of chemical references spaces and interpolation of compound potency surfaces (Peltason et al. 2010). To derive a 3D activity landscape the molecules are initially projected into a bi-dimensional chemical reference space that is spanned by two molecular descriptions defining the x and y axes. From a chosen molecular representation, a coordinate-free chemical reference space is generated by calculating pairwise compound distances. Then, multidimensional scaling (Borg and Groenen 2005) is used to project these molecules from the coordinate-free reference space onto an x/y -plane on the basis of their chemical distances. A third axis z is added accounting for the respective potency values of the molecules. Finally, a coherent potency surface required to obtain an interpretable landscape topology is obtained by interpolating data points by using a geostatistical technique termed Kriging (Cressie 1993). The interpolated potency surface is color-coded applying a gradient from green (lowest potency) to red (highest potency). A hypothetical exemplary representation of a three dimensional activity landscape is provided in the Fig. 3d. Following the same logic applied in the rest of ALM approaches the different SAR trends in a dataset can be intuitively identified by simply exploring the resultant landscape.

3.3 Are All Activity Landscapes Valid?

If all the approaches for ALM have proved to be valid, the same cannot be said for every activity landscape derived. In general, it is clear that both, activity landscapes

and activity cliffs, are very dependent on the context, which can be defined in multiple ways—by target, by chemical series or by molecular representation (Yongye et al. 2011; Medina-Franco 2012). The main factor contributing to such limitation is the strong dependence of activity landscapes of structure representation. Many are the structure representations currently available and in principle, any can be used to construct an activity landscape. However, not any structure representation provides a valid activity landscape, independently of the SAR dataset modeled. In the same way, not every activity cliffs derived from any structure representation can be considered as valid (statistically significant). The use of activity landscapes derived from a consensus of multiple structure representations (Yongye et al. 2011; Medina-Franco 2012) can alleviate this problem by identifying pairs of molecules that remains as activity cliffs in the multiple representations used (Medina-Franco et al. 2009). However, in absence of measures of significance, any approach to activity landscapes have to be simply assumed as valid and consequently, that a real SAR is present, which is not necessarily true.

In this direction, Guha and Medina-Franco (2014) alerted that methods that employ the landscape paradigm should first perform checks to assess the validity of the landscape. These authors proposed a method to validate the statistical significance of a given activity landscape which in essence answer the question of whether or not it is different from one generated by chance. For this, the authors resorted to the analogous application to activity landscaped of the concept of comparing predictive models to models developed using random (or scrambled) data, commonly used for QSAR models validation (Tropsha et al. 2003). The rationale behind this approach is that activity landscapes that emerge from structural representations not significantly different from a random landscape are “artifacts” and the structural representation does not reliably encode a real SAR (Medina-Franco 2013). That is, since similar landscapes can be generated from randomized activity and structural data, there is no true connection between the structural features captured by the representation and the activity data.

The same line of reasoning is applied to consider the validity of individual activity cliffs, once checked the validity of the landscape. For this, the authors used the SALI formalism by combining the idea of threshold-based identification of SALI cliffs (Guha and Van Drie 2008b; Guha 2011) with a permutation test to identify statistically significant activity cliffs. Depending on the ALM approach, other metrics such as SARI can be employed to quantify and detect activity cliffs pairs and consequently are susceptible to be combined with permutation tests to estimate the statistical significance of activity cliffs.

The results of the experiments presented in (Guha and Medina-Franco 2014) suggest that not all representations lead to non-random landscapes. This indicates that neither all molecular representations should be used to interpret the SAR nor be combined to generate consensus models. Consequently, significance testing of activity landscape and in particular, activity cliffs, is a key factor to consider prior to the application of ALM approaches.

3.4 Overview on Activity Landscape Modeling Applications

ALM is becoming a common strategy in medicinal chemistry and drug discovery to systematically explore and describe SAR trends. Although specific applications have been reported, ALM have been mainly devoted to SAR analysis focused on the role of activity cliffs in compound class-specific or target-specific collections. Current ALM approaches have provided the platform for identifying local SAR trends in large high throughput screening datasets (Wawer and Bajorath 2009); systematic studies of chemical substitutions (Wassermann and Bajorath 2010) or scaffolds (Hu and Bajorath 2010) that exhibit a propensity for activity cliff formation; structure—selectivity relationships (Peltason and Bajorath 2009); definition and identification of activity ridges (multiple compounds in a series forming activity cliffs) (Vogt et al. 2011); and the introduction of concepts such as ‘multitarget activity landscapes’ (derived from compounds that are active against multiple targets) (Dimova et al. 2011) or ‘R-cliffs’ (activity cliffs derived from a pair of compounds that differ in a single R group) (Agrafiotis et al. 2011). A detailed summary of ALM approaches to SAR analysis is provided in (Wassermann et al. 2010).

Among the most salient applications of the ALM approach, not directly dealing with SAR analysis, are their use in the estimation of predictive models performance and quality (Guha and Van Drie 2008a); the introduction of consensus approaches allowed to identify representation independent activity cliffs termed as ‘consensus activity cliffs’ (Medina-Franco et al. 2009) as well as ‘activity cliffs generators’ (ACG) (Méndez-Lucio et al. 2012b); feature selection approaches based on quantitative ALM metrics such as SALI, directed to find a structural representation that leads a to a good activity landscape and consequently to a predictive model (Guha 2012c); or activity landscape prediction approaches based on ligand structure (Guha 2010) or both ligand and protein information (Seebeck et al. 2011). A summarized overview of this applications is provided in (Guha 2012b).

More recent applications explore novel or emerging areas beyond biological activity. An attractive and non-common structure representations termed PLIFs (Brewerton 2008) was recently used to introduce the concept of ‘interaction cliffs’ (ligand-target complexes with high structural and interaction similarity but with a large potency difference between ligands) (Méndez-Lucio et al. 2015). This type of analysis constitutes a primary example of the synergy between chemoinformatics and molecular modeling to process efficiently large amounts of data from protein–ligand and protein–protein complexes (Medina-Franco 2014). A novel representation to address absorption, distribution, metabolism and elimination (ADME) properties and model property landscape models (PLM) is represented by the recent study of Yongye and Medina-Franco that discussed the structure–property relationships (SPR) of 166 molecules screened for kappa-opioid receptor activity including ADME considerations (Yongye and Medina-Franco 2013). As part of the emerging area of Food informatics (Martinez-Mayorga and Medina-Franco 2014), ALM was recently applied to mine structure–flavor associations. Finally, the problem of scaffold hopping (finding/predicting/screening compounds that are

structurally diverse, while sharing a biological activity) (Schneider et al. 1999; Tsunoyama et al. 2008) have been also addressed through the ALM approach leading to the introduction of the concept of ‘similarity cliffs’ (compounds in a compound-pair having approximately equal activities but significantly different structures) (Vogt et al. 2013). A detailed discussion on this novel applications can be found in (Medina-Franco et al. 2015a).

3.5 *Importance of Integrating ALM with QSAR*

QSAR modeling is one of the major computational tools employed in medicinal chemistry and drug discovery. However, from the beginning, there has been the focus of both approval and criticism concerning its reliability (Cherkasov et al. 2014). In essence, despite almost limitless availability of molecular descriptors and the increasing variety and efficiency of machine learning techniques available for implementing QSAR models, their predictive capability is still limited. Significant wrong predictions of activity still arise among similar molecules even in cases where overall predictivity is high. Unfortunately, this observation made by Maggiora (2006) still holds 10 years later. As noted in his editorial (Maggiora 2006), the reason why QSAR often disappoints is essentially related to the nature of the underlying SAR. That is, the main assumption of QSAR and similarity-based approaches is SAR continuity. As above mentioned, it is founded on the similarity property principle (Johnson and Maggiora 1990) and consequently assumes that gradual changes in structure should necessarily leads to gradual changes in activity. However, systematic quantitative profiling of many different sets of active compounds has shown that the majority of global SARs are heterogeneous in nature (Peltason and Bajorath 2007a), that is, their activity landscapes contain both gently sloped regions (continuous SAR) but also sharp cliffs (discontinuous SAR). Therefore, the presence of SAR continuity provides a fundamental basis for QSAR analyses and resulting compound activity predictions, while the presence of SAR discontinuity falls outside the applicability domain of the QSAR paradigm (Maggiora 2006).

Currently, machine learning algorithms are the most extended tools in QSAR modeling and chemoinformatics applications (Witten and Frank 2005; Ivanciuc 2009; Aguiar-Pulido et al. 2013). As discussed above, since QSAR modeling is essentially based on a similarity-based approach, the meaningful application of machine learning algorithms to chemoinformatics and drug discovery tasks heavily relies in the fulfilment of the Similarity Property Principle.

The two general purposes for which machine learning is used in QSAR modeling are classification and generalization of data, where machine learning is used to extract regularity from data. In a drug discovery context, machine learning uses SAR knowledge to guide the process in favor of producing classifications and generalizations that are conceptually meaningful (Rose 2003). Accordingly, special data cases such as activity cliffs represent exceptions for the regular trend encoded in the whole data; or even worst, represent examples that contradicts that regular trend.

Therefore, if the classification mechanism in machine learning is understood as a function that maps a description (chemical structure encoded by molecular descriptors) of an example to its label, i.e., a continuous value or a class membership, the counterproductive influence of pair of molecules that form activity cliffs is obvious. That is, most machine learning techniques just capture major trends encoded in continuous SAR regions but fail to recognize activity cliffs. Thus, due to the very likely variable constitution of an activity landscapes the reliability of the resultant predictions will be reduced. Even for advanced techniques capable of handling nonlinear relationships such as neural networks or support vector machines (SVMs), it is difficult to capture activity cliffs together with the rest of the landscape. But even if the machine learning model succeeds in capturing more of the significant activity cliffs it comes at a cost. To account for the most significant activity cliffs (which correspond to discontinuities in the SAR landscape), it would have to be “memorized” by the model, which introduces some degree of overfitting. Hence, a model that learns from a training dataset including a significant number of activity cliffs is prone to be overfitted (Guha 2011). Consequently, the conscientious application of SAR/SPR exploration based on ALM approaches previous to QSAR models development and deployment is a key to guarantee a meaningful decision making and a desirable prediction performance of the resultant models.

3.5.1 In Machine Learning Activity Cliffs Are Instances that Should Be Misclassified

The recently introduced concept of ‘instances that should be misclassified’ (ISMs) (Smith and Martinez 2011) can be used to establish a parallelism for activity cliffs in the machine learning area. In this paper, the authors introduced a novel filtering method that identifies ISMs using heuristics that predict how likely it is that an instance will be misclassified. By ISMs the authors mean that in the absence of additional information other than what the dataset provides, the label assigned by the learning algorithm to the instance is the most appropriate one, even if it happens to be different from the instance’s actual label. In this work ISMs are distinguished from traditional outliers and class noise on that ISMs exhibit a high degree of class overlap. That is, how similar an instance is to other instances of different classes in a region of the task space. A similar rationale is also behind the dataset ‘modelability index’ (MODI) proposed by Tropsha (Golbraikh et al. 2014) to estimate the feasibility of obtaining predictive QSAR models for a binary data set of bioactive compounds.

Although ISMs are defined in a classification context, a good analogy can be established with the activity cliff concept. If in machine learning terms ISMs are similar instances with different labels in a region of the task space, in activity landscape terms, activity cliffs represent small steps in chemical space that are accompanied by large changes in activity. However, this analogy becomes almost perfect if the activity cliff concept is extended to also take inactive compounds into consideration, reinterpreting the activity landscape as an active/inactive classification task rather than the usual regression task.

Smith and Martinez demonstrated that removing ISMs before training achieved the highest overall classification accuracy compared with the machine learning algorithms trained on the original data sets as well as with outliers removed by the other methods (Smith and Martinez 2011). Rather than focusing on correctly classifying the ISMs and arbitrarily adjusting the classification boundary, removing the ISMs for training allows the machine learning algorithms to focus on the instances that can be correctly classified. So, removing the ISMs allows a more appropriate decision surface to be discovered since the ISMs do not arbitrarily pull the decision surface from its more optimal position which leads to an improved classification accuracy. In terms of activity landscape, this can be translated into ‘remove activity cliffs to make the activity landscape smoother’.

Therefore, although in the QSAR and chemoinformatics community the benefits of removing problematic instances from training to improve the prediction accuracy of machine learning models is not widely accepted, from a machine learning perspective the effectiveness of this procedure is well justified and documented (Byeon et al. 2008; Smith and Martinez 2011; Smith et al. 2014; Yang and Gao 2013). Putting together the previous considerations give rise to the hypothesis: ‘removing activity cliffs will make the activity landscape smoother and improve the prediction accuracy of QSAR models’. After all, current QSAR and chemoinformatics tools are mainly based on machine learning algorithms.

3.5.2 SAR Continuity Restoration by Identification and Removal of Activity Cliffs Generators

To accept and keep in mind the negative influence of activity cliffs in QSAR modeling is the usual behavior among best QSAR practices (Scior et al. 2009; Tropsha 2010). Recently, the detection and identification of activity cliffs have been included as a key step in the process of chemogenomics data curation (Fourches et al. 2016). These authors propose that prior to initiating the computational study of a dataset, all activity cliffs forming compounds must be detected, verified, and treated by the modeler in order to decide whether to keep or discard them. However, little work has been devoted to alleviate it by reducing the SAR discontinuity on a dataset seeking to restore as much as possible the fundamental principle of QSAR and similarity-based methods.

In this sense, the closest concept to ISMs is ACG, which is defined in terms of single molecules instead of molecule pairs. An ACG is defined as a molecular structure that has a high probability to form activity cliffs with molecules tested in the same biological assay (Méndez-Lucio et al. 2012). So, problematic instances identified in machine learning terms as ISMs could be termed in chemoinformatics and computational medicinal chemistry terms as ACGs.

To minimize the risk of finding false ACGs, the reference space(s) used for their identification must be different and independent from the reference spaces that will be used for modeling. Ideally, the concept of consensus activity cliffs (Medina-Franco et al. 2009) (or a related concept such as the MMP-cliffs (Hu et al. 2012a, b)

should be applied for the identification of ACGs due to the well-known representation dependence of the activity cliffs concept (Dimova et al. 2012). In this way, the consensus ACGs identified by using several representations (preferably global representations such as fingerprints) should behave as such, irrespective of the reference space used or, at least, for most of the possible reference spaces.

Then, the goal is that the original training set, once removed the consensus ACGs previously identified, fulfills the assumption made by the machine learning algorithm intended to be used for model construction which is also the main premise of QSAR modeling: the SAR continuity implicit in the Similarity Property Principle. Additional curation (Fourches et al. 2010) and balancing procedures (Japkowicz 2000a, b) should be also applied to match the goals of both the QSAR paradigm (Maggiora 2006) and the machine learning algorithm (Witten and Frank 2005).

The essence of this solution is to remove from the training process those compounds responsible for the SAR discontinuity, and consequently restore the SAR continuity required for deriving reliable and predictive QSAR models. The main assumption behind this solution is that a machine learning algorithm that learns from a training set free of the noise induced by these problematic examples should produce a model able to identify the structural and/or physicochemical patterns determining the desired activity in a sharper way than when learning from a training set including those problematic examples. However, the question that remains is to what extent the learning process is affected and so, the generalization ability of the pattern found by the loss of the information encoded in the activity cliff pairs (Maggiora 2006). Other drawback of removing ACGs is the unavoidable reduction of the applicability domain of the model.

A remedial measure to soften the loss of applicability domain can be that of deriving several diverse machine learning models to implement a consensus classifier (Zhang et al. 2013; Polikar 2006). It is well known that multi-classifiers, ensemble or consensus classifiers are effective, among other reasons, because they span the decision space since each base classifier covers a different region of the decision space (chemical space or SAR) and the union of all the base classifiers produce a common region that results in a wider chemical coverage or applicability domain (Kuncheva 2004; Polikar 2006). So, in our opinion, at least it is worth to test the hypothesis of ACGs removal since reduction of the applicability domain seems to have a remedial solution whereas overfitting does not.

Actually, a successful example of the positive effect of removing ACGs in prospective virtual screening tasks was recently provided by Cruz-Monteaudo et al. (Castillo-Gonzalez et al. 2015). Here the authors proposed a virtual screening methodology for the discovery of novel G-quadruplex stabilizing agents with a hit rate higher than 20%. This methodology combined a consensus QSAR modeling and molecular docking including as a key step in the QSAR modeling stage the detection and removal of ACGs. However, even when the results obtained by the application of such VS methodology are encouraging, removing ACGs could significantly aid or not such results. Additionally, this little more than anecdotic

evidence does not demonstrate whether or not the elimination of ACGs is certainly beneficial, detrimental or useless in terms of generalization ability.

This last, and specifically, the hypothesis of SAR continuity restoration by ACGs removal was recently probed in a recent work (Cruz-Monteagudo et al. 2016). In this work, Cruz-Monteagudo et al. report the first attempt to study the effect of activity cliffs over the generalization ability of machine learning based QSAR classifiers, using as a case study a previously reported diverse and noisy dataset focused on drug induced liver injury (Fourches et al. 2010) and more than 40 ML classification algorithms. Here, the hypothesis of SAR continuity restoration by activity cliffs removal is tested as a potential solution to overcome such limitation. Based on a previously established parallelism between ACGs and ISMs (Cruz-Monteagudo et al. 2014), the classification performance of multiple machine learning classifiers as well as the consensus classifier derived from predictive classifiers obtained from training sets including or excluding ACGs was comparatively studied.

The influence of the removal of ACGs from the training set over the virtual screening performance was also studied for the respective consensus classifiers algorithms. In general terms, the removal of the ACGs from the training process slightly decreased the overall accuracy of the ML classifiers and multi-classifiers, improving their sensitivity (the weakest feature of ML classifiers trained with ACGs) but decreasing their specificity. Although these results did not support a positive effect of the removal of ACGs over the classification performance of ML classifiers, the “balancing effect” of ACG removal demonstrated to positively influence the virtual screening performance of multi-classifiers based on valid base ML classifiers. Specially, the early recognition ability was significantly favored after ACG removal. Finally, although the results presented and discussed in this work provided evidences supporting the positive effect of ACG removal, mainly for virtual screening applications, extensive benchmark studies including multiple SAR datasets are still required to go beyond anecdotic evidences.

4 QSAR and ALM Studies Applied to Epigenetic Targets

As discussed above, the large amount of SEA stored in compound databases requires the application of computational approaches to extract useful information that can help to understand the activity of compounds at the molecular level. Similarly, computer-based analysis of structure-epigenetic activity information can be used to predict the activity of tested compounds, i.e., to conduct virtual screening of compound databases. Indeed, both applications have been published for epigenetic compounds. In this sub-section is discussed progress on the computer-based analysis of SEA. The discussion is organized in major parts: analysis of QSAR and ALM.

4.1 QSAR in Epigenetics

A number of QSAR studies have been reported for compounds screened across epigenetic targets. Several of these studies have been recently collected and discussed Méndez-Lucio (Méndez-Lucio 2016). Table 3 summarizes additional examples of QSAR models developed for epigenetic targets (Maldonado-Rojas et al. 2015; Sharma et al. 2013; Silvestri et al. 2012; Wei et al. 2012; Noor et al. 2015; Choubey et al. 2016; Sun et al. 2016). Most of the published QSAR studies have been done for HDACs and inhibitors (Méndez-Lucio 2016). The number of studies for BRDS is scarce.

Maldonado-Rojas et al. reported a QSAR study of 800 compounds to discriminate active and inactive DNMT inhibitors from a data collection of natural products. The QSAR model gave rise to the identification of 447 compounds for later molecular docking, cluster analysis and biological evaluation. The QSAR method included a Linear Discriminant Analysis, a statistical tool to obtain the best equation that permits the separation of two regions (actives-inactives). Statistical parameters for the QSAR model (Table 3) showed high correlation coefficients, accuracy, and sensitivity rate values.

As mentioned at the beginning of this section, most of the QSAR studies for epigenetic targets have been developed for HDACs. The main function of HDACs is to remove acetyl groups from the acetylated lysines located in the histone tails. Consequently, the positive charge of the amino group in the lysines interacts with the negative charge in the DNA phosphates, promoting chromatin condensation. As a result, gene silencing is observed since the transcription machinery cannot access gene-promoter regions (Sterner and Berger 2000). Gene silencing product of HDACs activity can produce different cancer types. For this reason several research groups are trying to identify and develop HDAC inhibitors. HDACs are classified structurally and functionally into several classes, namely class I (HDACs subtypes 1–3 and 8), class II (HDACs subtypes 4–7, 9 and 10), class IV (HDAC subtype 11), and class III (Ruijter et al. 2003). The class III HDACs (also known as the sirtuins SIRT1–7) is zinc-dependent but is structurally distinct from the other classes and requires the cofactor NAD^+ for their deacetylase function.

The paper published by Sharma et al. (Table 3) did not take into account specific HDAC isoforms, generating a 2D-QSAR model to design novel HDAC inhibitors in general. The study used a data collection of 34 compounds derived from α -amino souberic acid generating models with multiple linear regression analysis. The statistics of the best models are shown in Table 3. Other validation analyses also were developed, like partial least squares ($r^2 = 0.88$ and $r^2_{(CV)} = 0.76$) and neural networking analysis ($R^2_{\text{Training}} = 0.86$ and $R^2_{\text{Test}} = 0.66$). All of them generated comparable results which prove that the model formed is sound and has good predictivity.

More specific studies were carried out by Silvestrini et al. and Wei et al. who emphasize specific isoforms, class I, II, IV and HDAC4-Class II respectively. In both studies, 3D-QSAR studies were developed. In the first study the authors used

Table 3 Examples of QSAR studies for compounds targeting epigenetic targets

Epigenetic target	Data set size	Method	Validation results	Application	Reference
DNMT	Training set: 32 Test set: 15	QSAR	$C_{\text{Training}}^{\text{a}} = 0.94$ $C_{\text{Test}}^{\text{a}} = 0.87$ $Q_{\text{Total}}^{\text{b}}(\text{Training}) = 96.9$ $Q_{\text{Total}}^{\text{b}}(\text{Test}) = 93.3$ Sensitivity rate $\text{Training} = 94.7$ $\text{Sensitivity rate}_{\text{Test}} = 88.9$	Discrimination of active/inactive molecules as DNMT is from 800 NPs from NatProd Collection	(Maldonado-Rojas et al. 2015)
HDAC	Training set: 26 Test set: 8	2D-QSAR	$r_{\text{Training}}^2 = 0.86$ $r_{\text{Training(CV)}}^2 = 0.81$ Std. error = 0.29	To design novel histone deacetylases inhibitors	(Sharma et al. 2013)
HDACclass I, II, IV	Training set: 94 Test set MTS ^c : 10 Test set CTS ^d : 6 Test set LTS ^e : 1	3D-QSAR (COMBINER) ^f	Multiprobe: electrostatic-desolvation (ELE + DRY): $r^2 = 0.80$ $q_{\text{LOO}}^2 = 0.76$ $\text{SDEP}_{\text{LOO}}^g = 0.81$	Permits quantification of structure—activity relationships through the electrostatic (Coulombic) and van der Waals interaction energies as well as additional parameters, such as solvation energy	(Silvestri et al. 2012)
HDAC4 Class II	Training set: 19 Test set: 15	(3D-QSAR) CoMFA	$q_{\text{Training LOO}}^2 = 0.76$ $q_{\text{Training}}^2 = 0.75$ $r_{\text{Training}}^2 = 1.0$ Std. Error = 0.04 $r_{\text{Test}}^2 = 0.80$	To improve prediction accuracy on some of the less active compounds often encountered in building CoMFA, CoMSIA, or pharmacophore models, by means of consensus 3D-QSAR	(Wei et al. 2012)
		(3D-QSAR) CoMSIA	$q_{\text{Training LOO}}^2 = 0.76$ $q_{\text{Training}}^2 = 0.76$ $r_{\text{Training}}^2 = 1.0$ Std. Error = 0.06 $r_{\text{Test}}^2 = 0.81$		
HDAC Class I	Training set: 10 Test set: 71	2D-QSAR	$r_{\text{Training}}^2 = 0.93$ Std. Error = 0.24	Virtual screening of compound libraries	(Noor et al. 2015)

(continued)

Table 3 (continued)

Epigenetic target	Data set size	Method	Validation results	Application	Reference
HDAC1 Class I	Training set: 32 Test set: 10	3D-QSAR	$r^2 = 0.977$ $q^2 = 0.801$ Externally validation: $r^2_{pred} = 0.929$ $r^2_{cv} = 0.850$	Virtual screening against libraries	(Choubey et al. 2016)
SIRT1	Training set: 346 Test set: 8	QSAR	AUC ^b = 0.86 RMSE ⁱ = 0.79 $r_{Pearson}$ correlation = 0.75 $\rho_{Spearman}$'s rank = 0.74	Virtual screening of databases	(Sun et al. 2016)

^aMatthew correlation coefficient^bAccuracy value^cModeled test set (inhibitory activity against several HDAC isoform)^dCrystal test set (two HDAC-8 inhibitor complexes and four bacterial HDAC homologues)^eLargazole test set (HDAC isoforms: HDAC-1, HDAC-2, HDAC-3, and HDAC-6-1)^fComparative binding energy^gStandard deviation error-of-prediction^hArea under the curveⁱRoot mean square error^jCorrelation coefficient^kCross validation coefficient

LOO Leave-one-out

the comparative binding energy approach—COMBINER-, a novel comprehensive tool for data mining that uses a series of receptor-ligand complexes to quantify interaction energies with molecular mechanics (Ortiz et al. 1995). Three different energy fields were used to compute the enzyme-ligand interactions, electrostatics-(ELE), steric-(STE) and desolvation-(DRY). The results indicated the ELE + DRY model was the best (Table 3). In the second study, 3D-QSAR models were built by aligning inhibitor series of compounds with docking of a protein receptor into the active site. Based on same structurally aligned sets, pharmacophore hypotheses were also generated. The best values of CoMFA and CoMSIA models (Table 3) and pharmacophore hypothesis led to the development of a c-3D-QSAR (consensus) model whose aim was to improve the prediction accuracy on some of the less active compounds.

Noor et al. and Choubey et al. (Table 3) published papers focused on QSARs studies of HDAC inhibitors of Class I (HDAC 1–3 and 8). Noor et al. generated a 2D-QSAR model employing a multiple linear regression technique of compounds with HDAC inhibitory activity. This model was used jointly with pharmacophore models for virtual screening of compounds libraries.

In a similar manner, Choubey et al. generated pharmacophore hypothesis and QSAR models. However, they developed only one 3D-QSAR model that was applied to HDAC1. The model was employed for virtual screening against compound libraries in order to identify novel scaffolds which can be experimentally validated to design future drug molecule, the statistical significance reveals the high predictive power of 3D-QSAR model (Table 3). Regression analysis was done by constructing partial least squares factors. Descriptors like hydrogen bond acceptor (A), hydrogen bond donor (D), one positively ionizable (P), negatively ionizable (N), aromatic ring (R) and hydrophobic group (H) were used for representing the structures of the training set.

The most recent study summarized in Table 3 was reported by Sun et al., and that is about one isoform of HDAC Class III, also known as SIRT1 or sirtuin1. The authors had the goal to construct QSAR models of SIRT1 ligands for virtual screening of 1444880 chemical structures collected from molecular databases. The QSAR model was carried out using an inductive logic programming to generate molecular models of SIRT1 inhibitors. Inductive logic programming represents a particular model as formal logics that would facilitate inductive reasoning among data (examples or facts), background knowledge (rules), and hypotheses.

4.2 ALM in Epigenetics

As reviewed above, several QSAR studies have been developed for epigenetic targets. However, studies towards the identification of ‘epigenetic activity cliffs’ and epigenetic scaffold/R-hops is still limited and represents an area of research for further development (Méndez-Lucio 2016).

4.2.1 DNA Methyltransferase Inhibitors

ALM has also been applied to epigenetic data sets, in particular for inhibitors of DNMTs (Naveja and Medina-Franco 2015a). One of the first published studies aimed to explain at the molecular level the large activity difference of compounds identified from high-throughput screening (Medina-Franco et al. 2014). The sulfonamide SW155246 (Fig. 4) forms two activity cliffs with structural related analogues such that the methylation or loss of the hydroxyl group is associated with a large decrease in the biological activity. Validated molecular docking and induced-fit docking protocols were applied to the sulfonamide compounds. It was concluded that induced-fit docking had a significant impact on the docking scores and the predicted binding modes as compared to docking where the structure of the protein is kept fixed. It was also concluded that the most active compound, SW155246 had a distinct binding mode as compared to the inactive sulfonamide analogues occupying part of the co-factor binding site as well as the catalytic site. Overall, the results of the computational simulations were in excellent agreement with the experimental activity (Medina-Franco et al. 2014).

More recently, Naveja et al. conducted a comprehensive ALM of 280 compounds screened against DNMT1 and available in ChEMBL (Naveja and Medina-Franco 2015b). In that work, the chemical space of the compound data set was explored first yielding to main type of structures, namely, nucleoside and non-nucleoside analogues. Since each type of chemical structures had different activity landscapes, ALMs were developed independently for each data set.

Based on the contents of ChEMBL available at the time of the study, it was found that nucleoside analogues presented a rough and heterogeneous landscape with the existence of deep activity cliffs (similar compounds with large potency

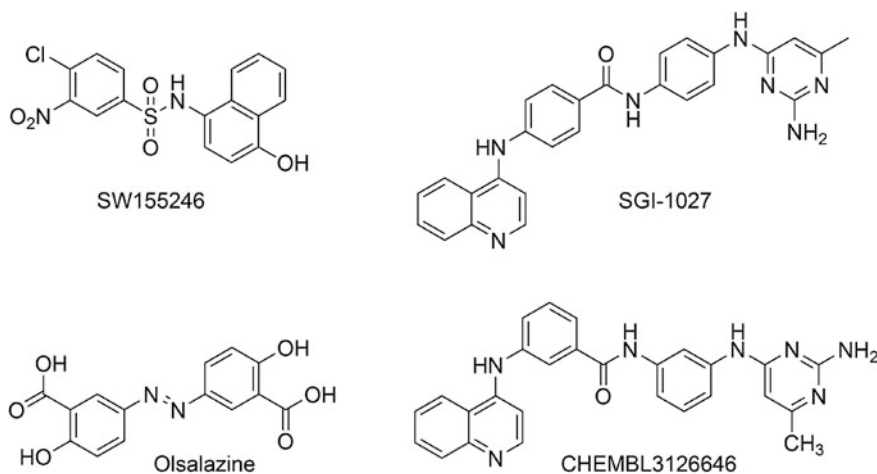


Fig. 4 Chemical structures of representative inhibitors of DNA methyltransferase or associated with DNA demethylating activity discussed in this work

difference). In contrast, non-nucleoside molecules presented a smoother SAR. The significance of that result was that almost any active non-SAM-like DNMT1 inhibitor in ChEMBL could be used as a reference in similarity-based virtual screening. Due to the presence of few activity cliffs, it was also concluded that non-nucleoside inhibitors represent a promising set of compounds to develop predictive models such as classical QSAR models. In contrast, the presence of deep activity cliffs for compounds related to the co-factor strongly suggest that that part of the chemical space may be not suitable to develop predictive QSAR models that rely on the similarity principle; e.g., classical predictive approaches that assume linear relationships (Naveja and Medina-Franco 2015b).

It was found that activity cliffs belong to the same structural class: regioisomers of the quinolone-based inhibitor SGI-1027 (Fig. 4). Molecular docking of the activity cliff generators with a crystallographic structure of DNMT1 further supported the idea that this type of inhibitors could be acting as stabilizers of the auto-inhibitory linker domain of DNMT1 (Yoo et al. 2013). The results of the docking model were also in line of the SAR of the activity cliffs formed with compound ChEMBL3126646 (Fig. 4). The results of the computational analysis were also in line with the experimental results that showed that ChEMBL3126646 is not a competitive inhibitor of the co-factor (Naveja and Medina-Franco 2015b).

As part of that work, density SAS maps (a modification to SAS maps) were employed to improve the visual interpretation and analysis (Naveja and Medina-Franco 2015a). Furthermore, the independent analysis of the activity landscape of compounds that are located indifferent regions of the chemical space led to the proposal of an ‘activity landscape sweeping’ approach: this is to explore systematically the SAR of structurally related compounds, e.g., explore local SARs. The ‘activity landscape sweeping’ methodology has been applied to conduct ALM of inhibitors of 5-alpha-reductase (Naveja et al. 2016) and can be extended to analyze the ALM of virtually any other epigenetic data sets.

4.2.2 Bromodomain Inhibitors

To further illustrate the application of ALM to the analysis of epigenetic data sets, it was conducted a preliminary survey of the activity landscape of a set of BRD inhibitors. A data set of 86 BRD subtype 4 (BRD4) inhibitors was considered as a case study. The set of compounds was obtained from ChEMBL and curated as recently reported (Prieto-Martinez et al. 2016). As discussed above, BRDs represent one of the major epigenetic targets not only to develop drugs but also molecular probes or chemical tools (Galdeano and Ciulli 2016).

Figure 5 depicts a SAS map for the 86 compounds. A SAS map portrays the relationship between the structure similarity (plotted on the X-axis) and the activity similarity or potency difference (plotted on the Y-axis) for all possible 3655 pairwise comparisons in the data set. In the map illustrated in the figure, the structural similarity was computed with MACCS keys fingerprints (Durant et al. 2002) and the Tanimoto coefficient (Jaccard 1901; Medina-Franco and Maggiora 2014). In

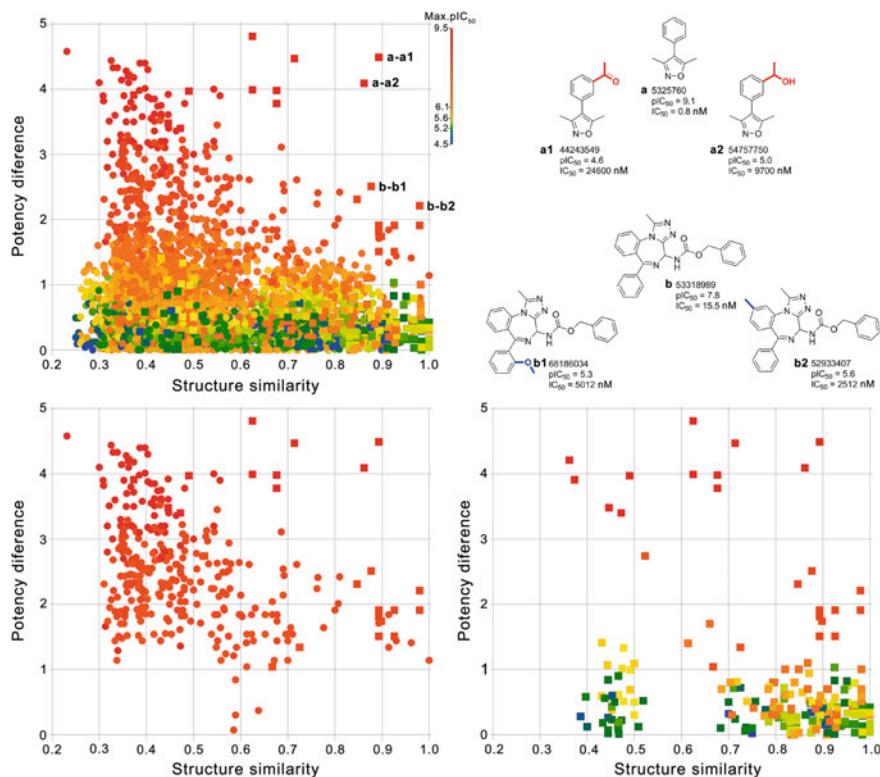


Fig. 5 Structure-Activity Similarity (SAS) map for 86 inhibitors of bromodomain 4. The plot in the upper part of the figure depicts 3655 pairwise comparisons. Data points are colored by the most active compound in the pair as indicated in the color scale. *Square data points* indicate pair of compounds with the same cyclic system. *Circle data points* indicate data points with different cyclic system. Four representative activity cliffs are labeled in the plot. SAS map on the lower left shows only active data points i.e., 334 pairs of molecules where the activity of the most active compound in the pair has a pIC₅₀ = 7.0 (IC₅₀ = 100 nM). SAS map on the lower right shows 226 data points where both compounds in the pair have the same cyclic system

order to identify the most interesting regions in the activity landscape, a SAS map can be roughly divided in four quadrants (Medina-Franco 2012). Although there are several considerations that should be taken into account to set up the thresholds to distinguish the four quadrants in a quantitative manner, as discussed above, there are at least two major regions in a SAS map of significant interest:

- The activity cliffs region located in the upper-right part of the graph contains pair of compounds with high structure similarity and high potency difference. In other words, this region of the SAS map identifies pair of molecules that have similar structures but their biological activity is significantly different, for instance, they differ in more than one logarithmic unit.

- The scaffold hop region located in the lower-left part of the graph, i.e., pair of compounds with low structure similarity and low potency difference. In other words, compounds that are quite different in structure but have similar biological activity.

Data points in the SAS map of Fig. 5 are further differentiated by a color that represents the activity of the most active compound in the pair. The color is in a continuous scale as indicated in the figure. The most attractive data points are those with at least one potent molecule in the pair. To facilitate the visualization of the most active pairs, Fig. 5 shows a SAS map showing only 334 pairs of compounds that contain at least one compound with $IC_{50} = 100$ nM ($pIC_{50} = 7.0$).

The SAS map in the case study shown in Fig. 5 also includes information of the cyclic system (also called molecular scaffold) of the molecules in each pair of compounds. The molecular scaffolds were computed with the program Molecular Equivalence Indices of Johnson and Xu (Xu and Johnson 2002). If both compounds in the pair share the same cyclic system, the data point has a squared shape. But if the cyclic system is different the shape of the data point is a circle. To facilitate the visual representation of the data points with the same or different cyclic systems, Fig. 5 includes a visualization of the SAS map showing only 226 data points where both compounds in the pair have the same cyclic system (square points). Similar strategies to represent pairs of compounds with the same or different cyclic systems in a SAS map have been reported (Pérez-Villanueva et al. 2012, 2015). This approach is valuable to add information of the cyclic system; information that is not directly encoded by the MACCS keys fingerprints. Interestingly and not surprisingly, compounds with the same cyclic system have, overall, higher values of molecular similarity than pairs of compounds with different cyclic system.

Examples of representative activity cliffs for inhibitors of BRD4 are labeled in Fig. 5. The pair of compounds with label **a-a1** (compounds with Compound Identifier, CID from PubChem: 44243549-5325760) and label **a-a2** (54757750-5325760) are the most pronounced, i.e., deep cliffs in the data set. These two pairs of molecules have MACCS keys/Tanimoto similarity of 0.893 and 0.862, respectively; and potency difference of more than four logarithm units i.e., 4.49 and 4.09, respectively. The most active compound in each pair (5325760) has an $IC_{50} = 0.8$ nM ($pIC_{50} = 9.1$). This compound is the most active in the entire data set. The chemical structures shown in Fig. 5 indicate that substitution of the phenyl ring of 5325760 has a pronounced effect in the activity decreasing the potency in about one thousand times. Both pairs of molecules **a-a1** and **a-a2** share the same cyclic system.

Two additional representative activity cliffs in data set of BRD4 inhibitors are the pairs of molecules labeled as **b-b1** (68186034-53318989) and **b-b2** (52933407-53318989) (Fig. 5). Also, chemical compounds in both pairs share the same cyclic system (as identified by the squares points) although the chemical scaffold is different from the pairs of molecules **a-a1** and **a-a2**. The most active compound in pairs **b-b1** and **b-b2** is molecule 53318989 which has an $IC_{50} = 15.5$ nM ($pIC_{50} = 7.81$). The position of the compounds pairs in the SAS map

indicate that pairs of molecules **b-b1** and **b-b2** have high structure similarity (0.877 and 0.980, respectively) but potency difference of more than two logarithmic units (2.51 and 2.21, respectively). Visual inspection of the chemical structures in the activity cliffs reveals that a methoxy (68186034) or methyl substitution (52933407) decreases the activity of 53318989 by more than hundred times. Further analysis of other data points and regions in the SAS map, including the scaffold hop region, can be performed as reported for several other data sets (Medina-Franco 2012).

5 Conclusions and Future Outlook

Epigenetic drug discovery is becoming a promising strategy to develop new therapies for several diseases. A number of computational approaches are being applied to speed up the development of epigenetic-based therapies. Elucidation of SEARS has been conducted with traditional and innovative approaches for a number of epigenetic targets including inhibitors of HDACs, DNMTs and BRDs, to name a few. Thus far, the principles of ALM have been applied mostly to inhibitors of DNMT and BRDs but they can be used in basically any other epigenetic data set to identify cliff generators or scaffold hops. ‘Activity landscape sweeping’ methodologies can be used to develop local models of SEARS. ALM can point to specific data sets suitable to develop predictive QSAR models and identify compounds to conduct similarity-based virtual screening. In line with the emerging concept of multi-epigenetic target drug discovery, structure-multiple epigenetic activity relationships can be performed. It is also anticipated that SEARS models will be used in drug repurposing. In fact, computer-aided drug repurposing has been applied to individual epigenetic targets such as DNMTs (Méndez-Lucio et al. 2014). However, it is expected that these strategies will be applied to other epigenetic targets.

Acknowledgements Valuable discussions with Oscar Méndez-Lucio, Eli Fernández-de Gortari, and J. Jesús Naveja are highly acknowledged. We also thank Fernando Prieto-Martínez for preparing the data set of bromodomain inhibitors. This work was supported by the Universidad Nacional Autónoma de México (UNAM), grant *Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica* (PAPIIT) IA204016 and *Programa de Apoyo a la Investigación y el Posgrado* (PAIP) 50009163, Facultad de Química, UNAM. MC-M acknowledges the postdoctoral grant [SFRH/BPD/90673/2012] financed by the FCT—Fundação para a Ciência e a Tecnologia, Portugal, co-financed by the European Social Fund.

References

- Agrafiotis, D. K., Wiener, J. J., Skalkin, A., & Kolpak, J. (2011). Single R-group polymorphisms (SRPs) and R-cliffs: An intuitive framework for analyzing and visualizing activity cliffs in a single analog series. *Journal of Chemical Information and Modeling*, 51, 31–1122.
- Agayo-Ortiz, R., Perez-Villanueva, J., Hernandez-Campos, A., Castillo, R., Meurice, N., & Medina-Franco, J. L. (2014). Chemoinformatic characterization of activity and selectivity switches of antiprotozoal compounds. *Future Medicinal Chemistry*, 6, 281–294.

- Aguiar-Pulido, V., Gestal, M., Cruz-Monteagudo, M., Rabunal, J. R., Dorado, J., & Munteanu, C. R. (2013). Evolutionary computation and QSAR research. *Current Computer-Aided Drug Design*, 9, 25–206.
- Alam, F., Islam, M. A., Gan, S. H., Mohamed, M. & Sasongko, T. H. (2016). DNA methylation: An epigenetic insight into type 2 diabetes mellitus. *Current Pharmaceutical Design*.
- Arguelles, A. O., Meruvu, S., Bowman, J. D., & Choudhury, M. (2016). Are epigenetic drugs for diabetes and obesity at our door step? *Drug Discovery Today*, 21, 499–509.
- Bajorath, J. (2012). Modeling of activity landscapes for drug discovery. *Expert Opinion on Drug Discovery*, 7, 463–473.
- Bajorath, J., Peltason, L., Wawer, M., Guha, R., Lajiness, M. S., & Van Drie, J. H. (2009). Navigating structure-activity landscapes. *Drug Discovery Today*, 14, 698–705.
- Berger, S. L., Kouzarides, T., Shiekhhattar, R., & Shilatifard, A. (2009). An operational definition of epigenetics. *Genes & Development*, 23, 781–783.
- Borg, I. & Groenen, P. J. F. (2005) *Modern Multidimensional Scaling. Theory and Applications*, New York, NY, Springer-Verlag.
- Brewerton, S. C. (2008). The use of protein-ligand interaction fingerprints in docking. *Current Opinion Drug Discovery Development*, 11, 356–364.
- Byeon, B., Rasheed, K. & Doshi, P. (2008) Enhancing the quality of noisy training data using a genetic algorithm and prototype selection. *The 2008 International Conference on Artificial Intelligence (ICAI'08)*. Monte Carlo Resort, Las Vegas, Nevada, USA: IEEE Publisher.
- Castillo-Gonzalez, D., Mergny, J. L., De Rache, A., Perez-Machado, G., Cabrera-Perez, M. A., Nicolotti, O., et al. (2015). Harmonization of QSAR best practices and molecular docking provides an efficient virtual screening tool for discovering new G-quadruplex ligands. *Journal of Chemical Information and Modeling*, 55, 110–2094.
- Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, Ii, Cronin, M., et al. (2014). QSAR modeling: Where have you been? Where are you going to? *Journal of Medicinal Chemistry*, 57, 4977–5010.
- Choubey, S. K., Mariadasse, R., Rajendran, S. & Jeyakanthan, J. (2016) Identification of novel histone deacetylase I inhibitors by combined pharmacophore modeling, 3D-QSAR analysis, in silico screening and density functional theory (DFT) approaches. *Journal of Molecular Structure*.
- Cressie, N. (1993). *Statistics for spatial data*. NY, New York: Wiley.
- Cruz-Monteagudo, M., Medina-Franco, J. L., Perez-Castillo, Y., Nicolotti, O., Cordeiro, M. N., & Borges, F. (2014). Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discovery Today*, 19, 80–1069.
- Cruz-Monteagudo, M., Medina-Franco, J. L., Perera-Sardina, Y., Borges, F., Tejera, E., Paz, Y. M. C. et al. (2016). Probing the hypothesis of SAR continuity restoration by the removal of activity cliffs generators in QSAR. *Current Pharmaceutical Design*.
- Desaphy, J., Raimbaud, E., Ducrot, P., & Rognan, D. (2013). Encoding protein-ligand interaction patterns in fingerprints and graphs. *Journal of Chemical Information and Modeling*, 53, 623–637.
- Dimova, D., Wawer, M., Wassermann, A. M., & Bajorath, J. (2011). Design of multitarget activity landscapes that capture hierarchical activity cliff distributions. *Journal of Chemical Information and Modeling*, 51, 258–266.
- Dimova, D., Hu, Y., & Bajorath, J. (2012). Matched molecular pair analysis of small molecule microarray data identifies promiscuity cliffs and reveals molecular origins of extreme compound promiscuity. *Journal of Medicinal Chemistry*, 55, 10220–10228.
- Dimova, D., Heikamp, K., Stumpfe, D., & Bajorath, J. (2013). Do medicinal chemists learn from activity cliffs? A systematic evaluation of cliff progression in evolving compound data sets. *Journal of Medicinal Chemistry*, 56, 45–3339.
- Dueñas-González, A., Jesús Naveja, J. & Medina-Franco, J. L. (2016). Chapter 1—Introduction of epigenetic targets in drug discovery and current status of epi-drugs and epi-probes. *Epi-informatics*. Boston: Academic Press.
- Durant, J. L., Leland, B. A., Henry, D. R., & Nourse, J. G. (2002). Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42, 1273–1280.

- Fernandez-de Gortari, E., & Medina-Franco, J. L. (2015). Epigenetic relevant chemical space: a cheminformatic characterization of inhibitors of DNA methyltransferases. *RSC Advances*, *5*, 87465–87476.
- Fourches, D., Barnes, J. C., Day, N. C., Bradley, P., Reed, J. Z., & Tropsha, A. (2010a). Cheminformatics analysis of assertions mined from literature that describe drug-induced liver injury in different species. *Chemical Research in Toxicology*, *23*, 83–171.
- Fourches, D., Muratov, E., & Tropsha, A. (2010b). Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of Chemical Information and Modeling*, *50*, 1189–1204.
- Fourches, D., Muratov, E. & Tropsha, A. (2016). Trust, but Verify II: A practical guide to chemogenomics data curation. *Journal of Chemical Information*.
- Galdeano, C. & Ciulli, A. (2016). Selectivity on-target of bromodomain chemical probes by structure-guided medicinal chemistry and chemical biology. *Future Medicinal Chemistry*.
- Golbraikh, A., Muratov, E., Fourches, D., & Tropsha, A. (2014). Data set modelability by QSAR. *Journal of Chemical Information and Modeling*, *54*, 1–4.
- Guha, R. (2010) What makes a good structure activity landscape? *Abstr Papers American Chemical Society*. Washington, DC: American Chemical Society.
- Guha, R. (2011). The ups and downs of structure-activity landscapes. *Methods Molecular Biology*, *672*, 101–117.
- Guha, R. (2012a). Exploring structure-activity data using the landscape paradigm. *Wiley Interdisciplinary Reviews: Computer Molecular Science*, *2*, 829–841.
- Guha, R. (2012b). Exploring structure-activity data using the landscape paradigm. *Wiley Interdisciplinary Reviews: Computer Molecular Science*, *2*, 829–841.
- Guha, R. (2012c). Exploring uncharted territories: Predicting activity cliffs in structure-activity landscapes. *Journal of Chemical Information and Modeling*, *52*, 2181–2191.
- Guha, R. & Medina-Franco, J. L. (2014) On the validity versus utility of activity landscapes: Are all activity cliffs statistically significant? *Journal of Cheminformatics* *6*, 11.
- Guha, R., & Van Drie, J. H. (2008a). Assessing how well a modeling protocol captures a structure-activity landscape. *Journal of Chemical Information and Modeling*, *48*, 1716–1728.
- Guha, R., & Van Drie, J. H. (2008b). Structure-Activity landscape index: identifying and quantifying activity cliffs. *Journal of Chemical Information and Modeling*, *48*, 58–646.
- Hu, Y., & Bajorath, J. (2010). Molecular scaffolds with high propensity to form multi-target activity cliffs. *Journal of Chemical Information and Modeling*, *50*, 500–510.
- Hu, Y., & Bajorath, J. (2012). Exploration of 3D activity cliffs on the basis of compound binding modes and comparison of 2D and 3D cliffs. *Journal of Chemical Information and Modeling*, *52*, 670–677.
- Hu, X., Hu, Y., Vogt, M., Stumpfe, D., & Bajorath, J. (2012a). MMP-cliffs: Systematic identification of activity cliffs on the basis of matched molecular pairs. *Journal of Chemical Information and Modeling*, *52*, 1138–1145.
- Hu, Y., Furtmann, N., Gütschow, M., & Bajorath, J. (2012b). Systematic identification and classification of three-dimensional activity cliffs. *Journal of Chemical Information and Modeling*, *52*, 1490–1498.
- Hu, Y., Stumpfe, D. & Bajorath, J. (2013). Advancing the activity cliff concept. *F1000Research*, *2*, 199.
- Huang, Z., Jiang, H., Liu, X., Chen, Y., Wong, J., Wang, Q., et al. (2012). HEMD: An integrated tool of human epigenetic enzymes and chemical modulators for therapeutics. *PLoS ONE*, *7*, e39917.
- Ivanciuc, O. (2009). Drug design with machine learning. In: R.A. Meyers (Ed.), *Encyclopedia of complexity and system science*. New York: Springer-Verlag.
- Iyer, P., & Bajorath, J. (2011). Representation of multi-target activity landscapes through target pair-based compound encoding in self-organizing maps. *Chemical Biology & Drug Design*, *78*, 739–905.
- Iyer, P., Hu, Y., & Bajorath, J. (2011). SAR monitoring of evolving compound data sets using activity landscapes. *Journal of Chemical Information and Modeling*, *51*, 40–532.

- Jaccard, P. (1901). Etude Comparative de la Distribution Florale dans une Portion des Alpes et des Jura. *Bull. Soc. Vaudoise Sci. Nat.* 37, 547–579.
- Japkowicz, N. (2000a). The class imbalance problem: Significance and strategies. *International Conference on Artificial Intelligence (ICAI 2000)*.
- Japkowicz, N. (2000b). Learning from imbalanced data sets: A comparison of various solutions. In R. Holte., N. Japkowicz, C. Ling & S. Matwin (Eds.), *AAAI 2000 workshop on learning from imbalanced data sets*. AAAI Press.
- Johnson, M. A., & Maggiora, G. M. (1990). *Concepts and applications of molecular similarity*. New York: Wiley.
- Kenny, P. W. & Sadowski, J. (2004). Structure modification in chemical databases. In T. I. Oprea (Ed.), *Cheminformatics in drug discovery*. Weinheim, Germany: Wiley-VCH.
- Kubinyi, H. (2002). From narcosis to hyperspace: The history of QSAR. *Quantitative Structure-Activity Relationships*, 21, 348–356.
- Kuncheva, L. I. (2004). *Combining pattern classifiers, methods and algorithms*. New York, NY: Wiley Interscience.
- Liu, T. Q., Lin, Y. M., Wen, X., Jorissen, R. N., & Gilson, M. K. (2007). BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Research*, 35, D198–D201.
- Liu, L., Zhen, X. T., Denton, E., Marsden, B. D., & Schapira, M. (2012). ChromoHub: A data hub for navigators of chromatin-mediated signalling. *Bioinformatics*, 28, 2205–2206.
- Lounkine, E., Wawer, M., Wassermann, A. M., & Bajorath, J. (2009). SARANEA: A freely available program to mine structure–activity and structure–selectivity relationship information in compound data sets. *Journal of Chemical Information and Modeling*, 50, 68–78.
- Maggiora, G. M. (2006). On outliers and activity cliffs—why QSAR often disappoints. *Journal of Chemical Information and Modeling*, 46, 1535.
- Maldonado-Rojas, W., Olivero-Verbel, J., & Marrero-Ponce, Y. (2015). Computational fishing of new dna methyltransferase inhibitors from natural products. *Journal of Molecular Graphics and Modelling*, 60, 43–54.
- Martinez-Mayorga, K., & Medina-Franco, J. L. (Eds.). (2014). *Foodinformatics: Applications of chemical information to food chemistry*. New York: Springer.
- Martinez-Mayorga, K., Peppard, T. L., Lopez-Vallejo, F., Yongye, A. B., & Medina-Franco, J. L. (2013). Systematic mining of generally recognized as safe (GRAS) flavor chemicals for bioactive compounds. *Journal of Agriculture and Food Chemistry*, 61, 7507–7514.
- Martinez-Mayorga, K., Peppard, L. T., Ramírez-Hernández, I. A., Terrazas-Álvarez, E. D. & Medina-Franco, J. L. (2014). Chemoinformatics analysis and structural similarity studies of food-related databases. In K. Martinez-Mayorga., L. J. Medina-Franco (Eds.), *Foodinformatics: Applications of chemical information to food chemistry*. New York: Springer.
- Medina-Franco, J. L. (2012). Scanning structure-activity relationships with structure-activity similarity and related maps: From consensus activity cliffs to selectivity switches. *Journal of Chemical Information and Modeling*, 52, 2485–2493.
- Medina-Franco, J. L. (2013). Activity cliffs: Facts or artifacts? *Chemical Biology and Drug Design*, 81, 553–556.
- Medina-Franco, J. L. & Maggiora, G. M. (2014). Molecular similarity analysis. In J. Bajorath (Ed.), *Cheminformatics for drug discovery*. Wiley.
- Medina-Franco, J. L., & Waddell, J. (2012). Towards the bioassay activity landscape modeling in compound databases. *Journal of the Mexican Chemical Society*, 56, 163–168.
- Medina-Franco, J. L. & Yoo, J. (2016). Chapter 15—The road ahead of the epi-informatics field. *Epi-Informatics*. Boston: Academic Press.
- Medina-Franco, J. L., Martinez-Mayorga, K., Bender, A., Marin, R. M., Giulianotti, M. A., Pinilla, C., et al. (2009). Characterization of activity landscapes using 2D and 3D similarity methods: Consensus activity cliffs. *Journal of Chemical Information and Modeling*, 49, 91–477.
- Medina-Franco, J. L., Edwards, B. S., Pinilla, C., Appel, J. R., Giulianotti, M. A., Santos, R. G., et al. (2013). Rapid scanning structure-activity relationships in combinatorial data sets:

- Identification of activity switches. *Journal of Chemical Information and Modeling*, *53*, 1475–1485.
- Medina-Franco, J. L., Méndez-Lucio, O., & Martínez-Mayorga, K. (2014a). Chapter one—The interplay between molecular modeling and chemoinformatics to characterize protein-ligand and protein-protein interactions landscapes for drug discovery. *Advances in Protein Chemistry and Structural Biology*, *96*, 1–37.
- Medina-Franco, J. L., Méndez-Lucio, O., & Yoo, J. (2014b). Rationalization of activity cliffs of a sulfonamide inhibitor of DNA methyltransferases with induced-fit docking. *International Journal of Molecular Sciences*, *15*, 3253–3261.
- Medina-Franco, J. L., Navarrete-Vázquez, G., & Méndez-Lucio, O. (2015a). Activity and property landscape modeling is at the interface of chemoinformatics and medicinal chemistry. *Future Medicinal Chemistry*, *7*, 211–1197.
- Medina-Franco, J. L., Navarrete-Vázquez, G., & Méndez-Lucio, O. (2015b). Property landscape modeling is at the interface of chemoinformatics and experimental sciences. *Future Medicinal Chemistry*, *7*, 1197–1211.
- Méndez-Lucio, O. (2016). Chapter 13—Computational structure–activity relationship studies of epigenetic target inhibitors. In J. L. Medina-Franco (Ed.), *Epi-Informatics*. Boston: Academic Press.
- Méndez-Lucio, O., Perez-Villanueva, J., Castillo, R., & Medina-Franco, J. L. (2012a). Activity landscape modeling of PPAR ligands with dual-activity difference maps. *Bioorganic and Medicinal Chemistry*, *20*, 32–3523.
- Méndez-Lucio, O., Perez-Villanueva, J., Castillo, R., & Medina-Franco, J. L. (2012b). Identifying activity cliff generators of PPAR ligands using SAS maps. *Molecular Informatics*, *31*, 837–846.
- Méndez-Lucio, O., Tran, J., Medina-Franco, J. L., Meurice, N., & Muller, M. (2014). Towards drug repurposing in epigenetics: Olsalazine as a novel hypomethylating compound active in a cellular context. *ChemMedChem*, *9*, 560–565.
- Méndez-Lucio, O., Kooistra, A. J., Graaf, C. D., Bender, A., & Medina-Franco, J. L. (2015). Analysing multitarget activity landscapes using protein-ligand interaction fingerprints: Interaction cliffs. *Journal of Chemical Information and Modeling*, *55*, 251–262.
- Naveja, J. J., & Medina-Franco, J. L. (2015a). Activity landscape of DNA methyltransferase inhibitors bridges chemoinformatics with epigenetic drug discovery. *Expert Opinion on Drug Discovery*, *10*, 1059–1070.
- Naveja, J. J., & Medina-Franco, J. L. (2015b). Activity landscape sweeping: insights into the mechanism of inhibition and optimization of DNMT1 inhibitors. *RSC Advances*, *5*, 63882–63895.
- Naveja, J. J., Cortés-Benítez, F., Bratoeff, E., & Medina-Franco, J. L. (2016). Activity landscape analysis of novel 5 α -reductase inhibitors. *Molecular Diversity*, *20*, 771–780.
- Noor, Z., Afzal, N., & Rashid, S. (2015). Exploration of novel inhibitors for class I histone deacetylase isoforms by QSAR modeling and molecular dynamics simulation assays. *PLoS ONE*, *10*, e0139588.
- Ortiz, A. R., Pisabarro, M. T., Gago, F., & Wade, R. C. (1995). Prediction of drug binding affinities by comparative binding energy analysis. *Journal of Medicinal Chemistry*, *38*, 2681–2691.
- Papadatos, G., & Overington, J. P. (2014). The ChEMBL database: a taster for medicinal chemists. *Future Medicinal Chemistry*, *6*, 361–364.
- Peltason, L., & Bajorath, J. (2007a). Molecular similarity analysis uncovers heterogeneous structure-activity relationships and variable activity landscapes. *Chemistry and Biology*, *14*, 489–497.
- Peltason, L., & Bajorath, J. (2007b). SAR index: Quantifying the nature of structure-activity relationships. *Journal of Medicinal Chemistry*, *50*, 5571–5578.
- Peltason, L., & Bajorath, J. (2009). Systematic computational analysis of structure activity relationships: Concepts, challenges and recent advances. *Future Medicinal Chemistry*, *1*, 451–466.

- Peltason, L., & Bajorath, J. (2011). Computational analysis of activity and selectivity cliffs. *Methods of Molecular Biology*, 672, 119–132.
- Peltason, L., Iyer, P., & Bajorath, J. (2010). Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. *Journal of Chemical Information and Modeling*, 50, 1021–1033.
- Perez-Villanueva, J., Santos, R., Hernandez-Campos, A., Giulianotti, M. A., Castillo, R., & Medina-Franco, J. L. (2010). Towards a systematic characterization of the antiprotozoal activity landscape of benzimidazole derivatives. *Bioorganic and Medicinal Chemistry*, 18, 91–7380.
- Pérez-Villanueva, J., Medina-Franco, J. L., Méndez-Lucio, O., Yoo, J., Soria-Arteche, O., Izquierdo, T., et al. (2012). CASE plots for the chemotype based activity and selectivity analysis: A CASE study of cyclooxygenase inhibitors. *Chemical Biology and Drug Design*, 80, 752–762.
- Pérez-Villanueva, J., Méndez-Lucio, O., Soria-Arteche, O., & Medina-Franco, J. (2015). Activity cliffs and activity cliff generators based on chemotype-related activity landscapes. *Molecular Diversity*, 19, 1021–1035.
- Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuit Systems and Magazine*, 6, 21–44.
- Prieto-Martinez, F. D., Gortari, E. F.-D., Mendez-Lucio, O. & Medina-Franco, J. L. (2016). A chemical space odyssey of inhibitors of histone deacetylases and bromodomains. *RSC Advances* 6, 56225–56239.
- Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50, 742–754.
- Rose, J. (2003). Methods for Data Analysis. In J. Gasteier (Ed.), *Handbook of Chemoinformatics*. Weinheim: Wiley-VCH.
- Ruijter, A. J. M. D., Gennip, A. H. V., Caron, H. N., Kemp, S., & Kuilenburg, A. B. P. V. (2003). Histone deacetylases (HDACs): Characterization of the classical HDAC family. *Biochemical Journal*, 370, 737–749.
- Schneider, G., Neidhart, W., Giller, T., & Schmid, G. (1999). Scaffold-hopping by topological pharmacophore search: A contribution to virtual screening. *Angewandte Chemie International Edition*, 38, 2894–2896.
- Scior, T., Medina-Franco, J. L., Do, Q. T., Martínez-Mayorga, K., Yunes Rojas, J. A., & Bernard, P. (2009). How to recognize and workaroud pitfalls in QSAR studies: A critical review. *Current Medicinal Chemistry*, 16, 4297–4313.
- Seebeck, B., Wagener, M., & Rarey, M. (2011). From activity cliffs to target-specific scoring models and pharmacophore hypotheses. *ChemMedChem*, 6, 1630–1639.
- Shanmugasundaram, V. & Maggiora, G. M. (2001). Characterizing property and activity landscapes using an information-theoretic approach. CINF-032. In *222nd ACS National Meeting, Chicago, IL, United States*. American Chemical Society, Washington, D.C.
- Sharma, S., Chauhan, R., Paliwal, S., & Dwivedi, J. (2013). 2-D QSAR model development for α -amino suberic acid derivatives as a novel anticancer agent. *Medicinal Chemistry Research*, 22, 1517–1527.
- Silvestri, L., Ballante, F., Mai, A., Marshall, G. R., & Ragno, R. (2012). Histone deacetylase inhibitors: Structure-based modeling and isoform-selectivity prediction. *Journal of Chemical Information and Modeling*, 52, 2215–2235.
- Smith, M. R. & Martinez, T. (2011) Improving classification accuracy by identifying and removing instances that should be misclassified. *The 2011 International Joint Conference on Neural Networks (IJCNN)*.
- Smith, M. R., Martinez, T., & Giraud-Carrier, C. (2014). An instance level analysis of data complexity. *Machine Learning*, 95, 225–256.
- Sterner, D. E., & Berger, S. L. (2000). Acetylation of histones and transcription-related factors. *Molecular Biology Review*, 64, 435–459.
- Stumpfe, D., & Bajorath, J. (2012a). Exploring activity cliffs in medicinal chemistry. *Journal of Medicinal Chemistry*, 55, 2932–2942.

- Stumpfe, D. & Bajorath, J. (2012b). Methods for SAR visualization. *RSC Advances* 2, 369–378.
- Stumpfe, D., Hu, Y., Dimova, D., & Bajorath, J. (2013). Recent progress in understanding activity cliffs and their utility in medicinal chemistry. *Journal of Medicinal Chemistry*, 57, 18–28.
- Sun, Y., Zhou, H., Zhu, H., & Leung, S.-W. (2016). Ligand-based virtual screening and inductive learning for identification of SIRT1 inhibitors in natural products. *Science Reports*, 6, 19312.
- Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics*, 29, 476–488.
- Tropsha, A., Gramatica, P., & Gombar, V. K. (2003). The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science*, 22, 69–77.
- Tsunoyama, K., Amini, A., Sternberg, M. J., & Muggleton, S. H. (2008). Scaffold hopping in drug discovery using inductive logic programming. *Journal of Chemical Information and Modeling*, 48, 57–949.
- Vogt, M., Wassermann, A. M., & Bajorath, J. (2010). Application of information—Theoretic concepts in chemoinformatics. *Information*, 1, 60–73.
- Vogt, M., Huang, Y., & Bajorath, J. (2011). From activity cliffs to activity ridges: Informative data structures for SAR analysis. *Journal of Chemical Information and Modeling*, 51, 1848–1856.
- Vogt, M., Iyer, P., Maggiora, G. M., & Bajorath, J. (2013). Conditional probabilities of activity landscape features for individual compounds. *Journal of Chemical Information and Modeling*, 53, 12–1602.
- Waddell, J., & Medina-Franco, J. L. (2012). Bioactivity landscape modeling: Chemoinformatic characterization of structure-activity relationships of compounds tested across multiple targets. *Bioorganic and Medicinal Chemistry*, 20, 5443–5452.
- Waddington, C. H. (2012). The epigenotype. *International Journal of Epidemiology*, 41, 10–13.
- Wassermann, A. M., & Bajorath, J. (2010). Chemical substitutions that introduce activity cliffs across different compound classes and biological targets. *Journal of Chemical Information and Modeling*, 50, 1248–1256.
- Wassermann, A. M., Wawer, M., & Bajorath, J. (2010). Activity landscape representations for structure-activity relationship analysis. *Journal of Medicinal Chemistry*, 53, 8209–8223.
- Wawer, M., & Bajorath, J. (2009). Systematic extraction of structure-activity relationship information from biological screening data. *ChemMedChem*, 4, 1431–1438.
- Wawer, M., & Bajorath, J. (2010). Similarity-potency trees: A method to search for SAR information in compound data sets and derive SAR rules. *Journal of Chemical Information and Modeling*, 50, 1395–1409.
- Wawer, M., Peltason, L., Weskamp, N., Teckentrup, A., & Bajorath, J. (2008). Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices. *Journal of Medicinal Chemistry*, 51, 6075–6084.
- Wei, H.-Y., Chen, G.-J., Chen, C.-L., & Lin, T.-H. (2012). Developing consensus 3D-QSAR and pharmacophore models for several beta-secretase, farnesyl transferase and histone deacetylase inhibitors. *Journal of Molecular Modeling*, 18, 675–692.
- Witten, I. H. & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. San Francisco, CA, Morgan Kaufmann.
- Xu, Y. J., & Johnson, M. (2002). Using molecular equivalence numbers to visually explore structural features that distinguish chemical libraries. *Journal of Chemical Information and Computer Sciences*, 42, 912–926.
- Yang, Z., & Gao, D. (2013). Classification for imbalanced and overlapping classes using outlier detection and sampling techniques. *Applied Mathematics and Information Science*, 7, 375–381.
- Yongye, A. B., & Medina-Franco, J. L. (2013). Systematic characterization of structure-activity relationships and ADMET compliance: A case study. *Drug Discovery Today*, 18, 732–739.
- Yongye, A., Byler, K., Santos, R., Martínez-Mayorga, K., Maggiora, G. M., & Medina-Franco, J. L. (2011). Consensus models of activity landscapes with multiple chemical, conformer and property representations. *Journal of Chemical Information and Modeling*, 51, 1259–1270.

- Yoo, J., Choi, S., & Medina-Franco, J. L. (2013). Molecular modeling studies of the novel inhibitors of DNA methyltransferases SGI-1027 and CBC12: Implications for the mechanism of inhibition of DNMTs. *PLoS ONE*, 8, e62152.
- Zhang, L., Fourches, D., Sedykh, A., Zhu, H., Golbraikh, A., Ekins, S., et al. (2013). Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. *Journal of Chemical Information and Modeling*, 53, 92–475.
- Zheng, Y. G. (2015). Chapter 1—The state of the art of epigenetic technologies. *Epigenetic technological applications*. Boston: Academic Press.

QSAR/QSPR Modeling in the Design of Drug Candidates with Balanced Pharmacodynamic and Pharmacokinetic Properties

George Lambrinidis, Fotios Tsopelas, Costas Giaginis
and Anna Tsantili-Kakoulidou

Abstract Drug discovery and development is a slow complicated multi-objective and expensive enterprise. Drug candidates are a compromise output of competing pharmacodynamics and pharmacokinetic processes. To facilitate this task and avoid failures in clinical phases, computational techniques and in silico modeling using the endpoints offered by high technology, are extremely valuable. In this chapter, some historical aspects and a background overview for constructing Quantitative Structure-Activity Relationships (QSAR) and Quantitative Structure-Property Relationships (QSPR) are provided. The different goals for the establishment of QSAR/QSPR models are defined. Representative examples and success stories of in silico modeling along the different drug discovery processes are presented. Examples include models for optimizing efficient binding to receptor, using both ligand- and structure-based approaches, for in vitro permeability predictions, predictions for human intestinal absorption and blood brain barrier penetration, as well as for plasma protein binding and drug metabolism. The value of global and local models as well as their interpretability and the criteria for their evaluation and proper use are discussed throughout this chapter.

G. Lambrinidis · A. Tsantili-Kakoulidou (✉)

Faculty of Pharmacy, Department of Pharmaceutical Chemistry, National and Kapodistrian University of Athens, Panepistimiopolis, 157 71 Athens, Zografou, Greece
e-mail: tsantili@pharm.uoa.gr

G. Lambrinidis

e-mail: lambrinidis@pharm.uoa.gr

F. Tsopelas

Laboratory of Inorganic and Analytical Chemistry, School of Chemical Engineering, National Technical University of Athens, Athens, Greece
e-mail: ftsop@central.ntua.gr

C. Giaginis

Department of Food Science and Nutrition, School of Environment, University of the Aegean, Limnos, Greece
e-mail: cgiaginis@aegean.gr

© Springer International Publishing AG 2017

K. Roy (ed.), *Advances in QSAR Modeling*, Challenges and Advances

in Computational Chemistry and Physics 24, DOI 10.1007/978-3-319-56850-8_9

Keywords Multi-objective drug discovery • Local models • Global models • Model interpretability • Drug-likeness • Affinity predictions • ADME properties predictions

1 Introduction

The advancement of a new chemical entity (NCEs) to become a drug candidate is a slow, complex, expensive and multi task process. Along this long road, identification of the disease and the isolation and validation of the molecular target(s) are the first crucial steps. Next, the right drug candidates to interact with the validated target are designed, synthesized and tested for their preclinical and clinical efficacy and safety (Satyanarayanajois 2011; Speck-Planche and Cordeiro 2015). Despite the great advances in science and technology, this process can take around 15 years with a cost of hundreds of millions of dollars (Paul et al. 2010). In fact, much of this cost comes from failures, which account for 75% of the total drug discovery and development expenses. On the other hand such failures if appropriately consolidated, contribute to the body of knowledge on biological complexity.

To prevent late-stage project interruptions, research is shifted to reduce the uncertainties and obtain a proof of concept (POC) for a molecule as a potential medicine in earlier phases of development. Thus, investigation of the fate of a molecule in the organism, considering appropriate pharmacokinetics as well as safety and adverse reactions profiles should advance in parallel with affinity for the target receptor(s) (Gaviraghi et al. 2001; Swift and Amaro 2013). The fate of drug molecules within the organism is principally controlled by ADME properties which stand for absorption, distribution, metabolism and elimination. (Rogge and Taft 2010; Testa et al. 2005b). Poor absorption and thereupon poor bioavailability have been in the past one of the main reasons for the failure of drug candidates. According to more recent statistics, the most important issues to be confronted are drug efficacy and drug safety, associated mainly with plasma protein binding, metabolism and off target activity (Kola and Landis 2004).

Computer-aided approaches and chemoinformatics, applied during the different stages of the pipeline, permit an effective handling of such failures and uncertainties, facilitate candidate selection and speed up their long journey to the market. Reliable models obtained by Quantitative Structure-Activity Relationships (QSAR) and Quantitative Structure-Property Relationships (QSPR) offer decision support upon rationalizing the drug discovery procedure in line with the Quick Win, Fast Fail concept, allowing a pre-selection of compounds with more chances to succeed in later phases (Owens et al. 2015). In this context, a new scientific area has emerged, defined as pharmacoinformatics, which enables the management of all available information from binding to kinetics and toxicity for safer drug candidates (Goldmann et al. 2014).

In fact, successful drug candidates usually represent a compromise between the numerous, sometimes competing objectives so that the advantages for patients

outweigh potential drawbacks and risks. However, in order to benefit from QSAR/QSPR models, the appropriate criteria for their evaluation and thereupon their proper use and/or interpretation are essential. Such criteria as well as the ultimate goal of the models may differ according to the timeline and the particular process modeled.

The present chapter provides an outline of the philosophy, the state of the art and the strategies for QSAR/QSPR generation. Distinction between QSAR and QSPR is primarily associated with the traditional drug design steps, concerning lead optimization for efficient receptor binding and predictions of pharmacokinetic/toxicity properties, respectively. After an overview of the common features for *in silico* modeling, QSAR models for pharmacodynamics properties, e.g., binding to target receptor(s) or off-target proteins and QSPR models for pharmacokinetic process (ADME properties) are discussed in separate sections. According to the underlying mechanism QSPR models concern both models for passive phenomena and for bonding to proteins. In all cases, two critical interdependent issues are addressed throughout the chapter: (i) the value of global models built on large and chemically diverse datasets and that of local models, built specifically for a series or project, and (ii) the importance or not of model interpretability (Cox et al. 2013; Fujita and Winkler 2016).

2 Historical Aspects and Background

Early QSAR studies were based on the assumption that biological activity can be quantitatively expressed as a function of chemical structure (Brown and Fraser 1868). They involved the establishment of model equations in order to understand and if possible to predict biological activity on the basis of structural parameters, as expressed by equation of type (1).

$$\text{Biological activity} = a_0 + a_1P_1 + a_2P_2 + \dots + a_nP_n \quad (1)$$

where $P_1 \dots P_n$ are physicochemical/molecular properties characterizing the compound structures and $a_0 a_1 \dots a_n$ the constants derived by multiple linear regression analysis (Hansch et al. 1995b; Hansch and Fujita 1964; Martin 1978).

Although biological activity was not always considered at the molecular level, it was recognized as an essential prerequisite that the analyzing compounds should act at the same receptor and with the same mechanism of action. Within a congeneric series it was assumed that all other factors influencing the manifestation of drug action should have similar impact. In regard to the description of chemical structure, the well-known Hansch analysis recognized three major categories of physicochemical parameters, namely lipophilicity, electronic properties and steric (geometric) properties (Eq. 2).

$$\log BR = -a \log P^2 + b \log P + \rho \sigma + \delta E_{\zeta} + c \quad (2)$$

where $\log BR$ is a general expression for biological activity in its logarithmic form to be linearly related to free energy, $\log P$ is the logarithm of octanol-water partition coefficient, the widely accepted measure of lipophilicity, σ Hammett's electronic substituent constant and E_{ζ} Taft's steric substituent constant (Hansch 1969; Hansch and Fujita 1964).

Evidently, early QSAR models could be developed only for congeneric compounds, having a common skeleton and different substituents. In those models, lipophilicity was considered as the physicochemical property of primary importance, since it was understood to influence both pharmacokinetics and pharmacodynamics (Kubinyi 1979; Leo et al. 1971; Pliška et al. 1996; Van de Waterbeemd and Testa 1987). A parabolic relationship between lipophilicity and membrane passage was assumed; thus the quadratic term in Eq. 2 reflects transport to the active site, considering all other pharmacokinetic issues equal within a congeneric series (Hansch and Clayton 1973). Since, the parabolic relationship between potency and $\log P$ did not fit all datasets, Kubinyi proposed a bilinear relationship, which allows for different slopes at low and high $\log P$ values (Kubinyi and Kehrhan 1978). At the same time calculation methods for $\log P$ were developed, based on the additivity principle. The hydrophobic substituent constant π and soon later the hydrophobic fragmental constant f or their $\Sigma\pi$ and Σf , accounting for all substituents/fragments on the parent structure, could replace $\log P$ of the whole molecule, in line with the other substituent constants in Hansch analysis (Hansch and Leo 1979; Rekker and Mannhold 1992).

In fact, Hansch analysis, firstly applied in agrochemistry, drug design, toxicology, industrial and environmental chemistry (Dunn 1988; Hansch et al. 1995a, 1963; Muir et al. 1967), marked a breakthrough in the way of thinking in medicinal chemistry and the start of the new discipline of QSAR (Ganellin 2004), with the mission to exploit the increasing amount of information in the aim to facilitate drug discovery.

Since those early days, QSAR has undergone a tremendous evolution in regard to all aspects, the target end points, the structural representation, the implemented statistical tools, as well as its own standpoints (Cherkasov et al. 2014; Cramer 2012; Puzyn et al. 2010; Tsantili-Kakoulidou and Agrafiotis 2011). In view of biological complexity QSAR has adapted to the multi-task concept, taking advantage of technological achievements and moving from the perception of single-objective drug design to the multi-objective drug discovery and development (Fujita and Winkler 2016; Jorgensen 2004; Speck-Planche and Cordeiro 2015). The multiple tasks addressed by QSAR/QSPR and the tools implemented to construct the models are illustrated in Fig. 1.

Thus, QSAR/QSPR models are generated to address two goals, each of which has its own value: One goal is to establish models which provide an insight of the properties or chemical features that correlate with a biological assay and thereupon an understanding of the mechanism of action. Such models are valuable support for

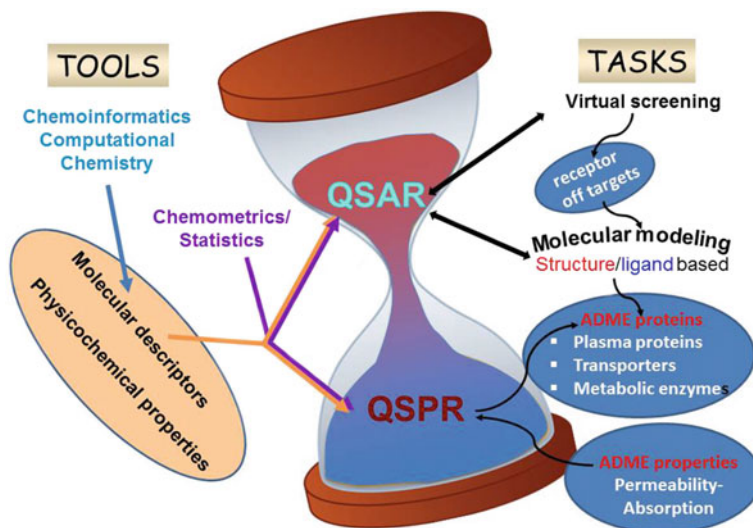


Fig. 1 Tasks addressed by QSAR/QSPR and tools implemented in model construction

the design of novel compounds with affinity to a target protein. The second goal is to create models, which provide accurate prediction of large chemically diverse datasets and address a variety of biological endpoints, as well as different pharmacokinetic processes. Such models allow ranking of compounds prior to synthesis or set priorities among drug candidates for proceeding to further development (Birchall et al. 2008a, b; Nicolotti et al. 2002).

3 Experimental Data and Endpoints in QSAR/QSPR

The multi-objective QSAR starts with data analysis for hit identification, followed by hit-to-lead optimization (lead discovery) and lead optimization (Jorgensen 2009). For hit identification, virtual screening has gained a crucial role, as a consequence also of the continuous emergence of novel biological targets (Schneider 2010; Vasudevan and Churchill 2009). QSAR end-points are usually measured at the molecular or cellular level. The advent of robotized biological testing in the 1990s (Ashour et al. 1987; Houston and Banks 1997; Löfås and Johnsson 1990; Navratilova et al. 2007) has led to the creation of large databases, freely accessible in the public domain, which incorporate millions of compounds with associated bioactivities. PubChem (<https://pubchem.ncbi.nlm.nih.gov>) and ChemSpider (<http://www.chemspider.com>), the two major collections of chemical structures on the web, currently include over 30 million compounds each. ZINC (<http://zinc.docking.org>), a database frequently used for virtual screening applications, incorporates a total of approximately 21 million compounds (Irwin 2008; Moura Barbosa

and Del Rio 2012; Wang et al. 2012). In such databases the results of many screens are presented in the form of scores for many compounds on a given assay, while they also contain information on the structures of compounds and the target of particular assays. More detailed data about binding assays can also be found in Binding Database (www.bindingdb.org) which is a public web-accessible database of measured binding affinities containing more than 1 million binding data for nearly 500,000 small molecules and thousands of proteins (Gilson et al. 2016).

However there is a warning on the use of the databases, since they may include inconsistencies concerning both chemical and biological data, while the chemical structures may be inaccurate or presented in a non-consistent way. Therefore curation of the data sets is recognized as a critical step for the establishment of good quality models (Akhondi et al. 2012; Cherkasov et al. 2014).

More to the point, there are databases with sets of inactive compounds (decoys) for several biological targets together with a small set of known active compounds (Mysinger et al. 2012) or even software to produce decoy datasets based on similarity with known active compounds (Cereto-Massagué et al. 2012). Decoy data sets are useful for validation of the QSAR/QSPR models.

When searching in structural databases for experimental binding affinities, one could find different biological data. They may be expressed as continuous response such as IC_{50} , EC_{50} , K_i , K_d , % inhibition, etc., or as categorical response, e.g., active/inactive. Continuous response values are preferably used in their negative logarithms, so as to be in linear correlation with free energy. In line with this concept, ChEMBL database introduced the pChEMBL activity value, defined as $-\log(IC_{50}, XC_{50}, EC_{50}, AC_{50}, K_i, K_d \text{ or Potency})$ in M units (Papadatos et al. 2015). This value allows a number of roughly comparable measures of half-maximal response concentration/potency/affinity to be compared on a negative logarithmic scale (<https://www.ebi.ac.uk/chembl/faq#faq67>). This approach has also been implemented in software for large scale off-target pharmacology and predictive safety of small molecule such as CTLink (<http://www.chemotargets.com>).

Besides the compound databases, there is also a wealth of deposited gene expression data available for downloading and/or online interrogation. For example, the NCBI gene expression omnibus (GEO) (Barrett et al. 2007) hosts over half a million single array chip expression profiles and the EBI hosts the Array Express database (Parkinson et al. 2010) with a similar largely overlapping number of arrays. Gene expression-based screening (GE-HTS) represents a strategy for identifying modulators of biological processes with little a priori information about their underlying mechanisms. It is mainly used in cancer research, where it detects compounds, which may revert undesired oncogenic states to nonmalignant or drug-sensitive states (Evans and Guy 2004; Williams 2012). It is evident that for the screening procedure, good prediction models are necessary, complying with the second goal as described in Sect. 2. In such case model interpretability is not a priority. In contrast, the transition from hit identification to lead discovery and optimization requires models which should provide an understanding of the molecular factors involved and a sound physicochemical interpretation, while in-house affinity measurements of the novel compounds are used as endpoints.

The range of affinity values is a crucial issue for model construction. Generally it should be significantly greater than the experimental error among the biological data. Considering that such errors can often exceed half a log unit (Gedeck et al. 2006) it is recommended an endpoint value range of at least 1.0 log unit to obtain a reasonable QSAR model (Cherkasov et al. 2014).

Lead optimization in regard to other pharmaceutical properties, while maintaining affinity, is a next important step. This is a multi-objective process involving many experimental parameters (assays) related to physicochemical properties, ADME properties, plasma and tissue protein binding, target selectivity, off-target activities and toxicity. These properties influence considerably the efficacy and safety of drug candidates and are potential causes for attrition. Rapid in vitro measurements have been and are being developed for permeability and for plasma protein binding assessment and toxicity protocols have been established (Artursson et al. 2001; Kansy et al. 1998; Kariv et al. 2001; Rich and Myszka 2000). On the other hand, there are many efforts for in silico prediction of many of these endpoints by constructing appropriate QSARs or QSPRs (A Cabrera-Perez et al. 2012; Dearden 2007; Lambrinidis et al. 2015; Swift and Amaro 2013). Certain global models for toxicity predictions are approved by OECD and provide support to regulatory authorities (Larregieu and Benet 2013). More to the point, predictions on secondary targets may be useful for the safety profile as well as for drug repurposing (Hodos et al. 2016; Sheridan et al. 2015).The implementation of QSAR/QSPR in the complex drug discovery process is demonstrated in Fig. 2.

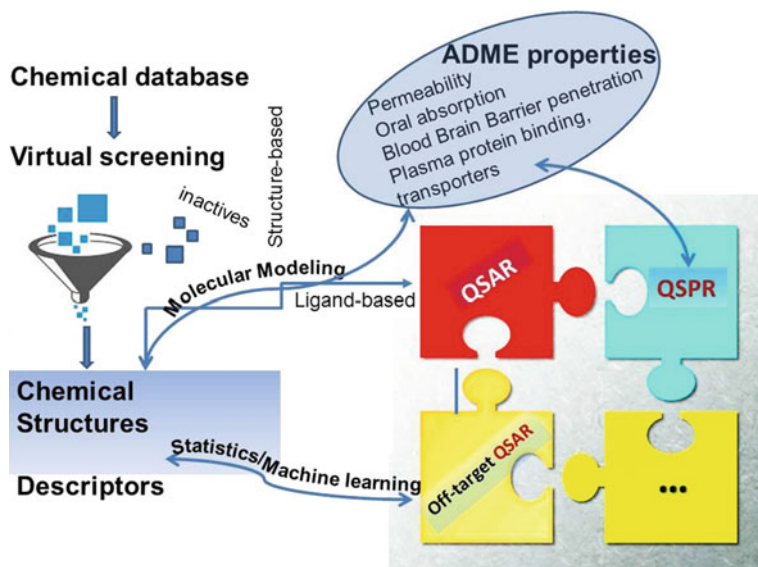


Fig. 2 Implementation of QSAR/QSPR in the drug discovery process

The splitting of QSAR models to encompass various areas of biological complexity has challenged the development of workflows, which integrate QSAR/QSPR models of selected endpoints, including affinities for different target proteins/off-targets and pharmacokinetic data (Cartmell et al. 2005). Consensus predictions using all acceptable models may contribute to further decisions in selecting future experimental screening sets. In inductive knowledge transfer approaches, treating multi-task modeling, the individual QSAR models are not considered separately but they are viewed as nodes in a network of inter-related models (Cherkasov et al. 2014; Qiu et al. 2016). Evidently, the quality of such integrated models largely depends on the quality of the available experimental data compiled in relevant databases, which should be carefully curated, as well as on the range of endpoint values, as already commented (Cherkasov et al. 2014; Gedeck et al. 2006). Interpretability of such models as a prerequisite depends on the purpose and the timeline that they are used along the drug discovery process. In regard to toxicity, for QSAR models to be accepted for regulatory purposes, interpretability is often a crucial issue. According to OECD *“To facilitate the consideration of a QSAR model for regulatory purposes, it should be associated with ... a mechanistic interpretation, if possible”* (www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf).

4 Tools Implemented in Model Construction

4.1 Molecular Structure Representation-Descriptors

Molecular structures are represented by descriptors which mediate their relation with activity. Thus, molecular descriptors are at the core of QSAR modeling.

In line with the definition of Todeschini and Consonni (2009), molecular representation has moved forward from substituent constants to variables suitable to portray diverse molecules, belonging to different chemical classes. A variety of software calculates a large number of different physicochemical/molecular properties and theoretical descriptors, starting from SMILES, 2D-chemical graphs to 3D-x, y, z-coordinates or based on mathematical algorithms or statistics. Some of the most popular software are DRAGON, which calculates more than 4000 descriptors (http://www.taletе.mi.it/products/dragon_description.htm), ADAPT (Stuper and Jurs 1976) (<http://research.chem.psu.edu/pcjgroup/adapt.html>), OASIS (Mekenyan and Bonchev 1986), CODESSA (Katritzky et al. 1994), MOE-Chemical Computing Group (<https://www.chemcomp.com/>) and MolConnZ (<http://www.edusoft-lc.com/molconn/>).

According to molecular structure representation, descriptors may reflect various levels of dimensionality, ranging from 0D to 4D and xD. 0D are based on molecular formula and are independent from molecular connectivity and conformations. 1D descriptors, reflect the substructure representation of a molecule, 2D descriptors are based on the two-dimensional structural formula (2D), while 3D descriptors are

conformation dependent. 3D descriptors are based on thermodynamically favored conformation and necessitate geometry optimization. 4D descriptors reflect interactions with some probe within a grid, while higher dimension (xD) are receptor dependent descriptors. They represent each ligand molecule as an ensemble of conformations, orientations, tautomeric forms and protonation states (Ekins et al. 1999; Hopfinger et al. 1997; Vedani et al. 2000, 2005). Using enhanced molecular dynamic simulations, the overall conformational change of the receptor upon ligand binding can be simulated, producing more vital structural descriptors (Sohn et al. 2013). Such approaches can be considered as a promising link between structure and ligand based strategies (Polanski 2009; Caporuscio and Tafi 2011). An atlas of the available descriptors, the theory used for their calculation and their information content, has been compiled by Todeschini and Consonni (2009). In Table 1, a classification of representative descriptors is presented.

Among the physicochemical descriptors, logP keeps its central role in drug-protein and drug-membrane interactions, as well as in permeability models. Nowadays, there are many algorithms for logP or logD calculation, implemented in relevant software. They are based on the additivity principle and have been developed upon analysis of a large amount of experimental data (Mannhold and Dross 1996). More to the point, calculation of logD necessitates knowledge on pK_a , while charge is a crucial determinant also in drug action (Csizmadia et al. 1997). Actually most of the logP, pK_a and solubility prediction algorithms are QSPR models per se. Some global logP models are implemented in software workflows, which allow the user to utilize his/her own compound library as input in order to refine predictions (Tetko et al. 2001). A comprehensive description and classification of the logP/logD calculation systems and software is provided by Mannhold et al. (Mannhold et al. 2009). Among them, ClogP is often considered as a reference calculation system, while it has been included in most rules for druglikeness (see Sect. 5). Some software for logP/logD prediction are free available on the web.

Despite the large arsenal of available software, the correct selection for logP/logD prediction is not always easy, since often the outcome of the different algorithms shows considerable variations. Although this is not an issue for models intended to screen large compound libraries, it becomes crucial for local models established for lead optimization or for predictions within congeneric compounds (Chrysanthakopoulos et al. 2009; de Melo et al. 2009). In such cases it is important that the compounds analyzed fall within the applicability domain of the training set, used to construct the prediction algorithm (see Sect. 4.3) (Tetko et al. 2009).

Next to lipophilicity, other molecular properties such as molecular volume and surface area, polarizability, molar refractivity, polarity descriptors, dipole moments, hydrogen bond acidity/basicity, as well quantum chemical descriptors, including energy parameters like E_{HOMO} and E_{LUMO} , maximum and minimum electrostatic potentials, partial charges etc., are most commonly used by medicinal chemists. Such descriptors considered as “well-founded”; actually fall within the frame of the three categories: lipophilicity, electronic and geometric descriptors, reflecting the recognition forces and steric requirements of binding to receptor active site.

Table 1 Classification of representative descriptors according to the theory used

Information content/ Theory used	Representative descriptors	Dimensionality	Representative publications
Physical Chemistry	logP solubility ionization constants polarizability molar refractivity ...	2D 2D 2D 2D 2D	Hansch et al. 1995b
Molecular properties	molecular weight molecular volume molecular surface hydrogen bond descriptors polarity flexibility-rotatable bonds angles distances ...	0D 3D 3D 2D 2D 2D 3D 3D	Helguera et al. 2008
Linear Solvation Energy	hydrogen bond acidity hydrogen bond basicity dipolarity/polarizability excess molar refractivity Mac Goyan Volume	2D 2D 2D 2D 2D	Abraham et al. 2002
Quantum Chemistry	partial charges energy parameters electrostatic potentials dipole moment electron density descriptors HOMO-LUMO energy gap ...	3D 3D 3D 3D 3D 3D	De Benedetti and Fanelli 2010

(continued)

Table 1 (continued)

Information content/ Theory used	Representative descriptors	Dimensionality	Representative publications
Molecular Interaction Field, MIF*	GRID descriptors CoMFA CoMSIA	4D 4D 4D	Goodford 1985; Cramer et al. 1988; Klebe et al. 1994
Molecular Interaction Fields (MIF) projected at different energy levels	Volsurf**	2.5D	Cruciani et al. 2000
Representation of molecules as ensemble of conformations, orientations, tautomeric forms, protonation states or any other experimental condition		xD	Vedani et al. 2005
Constitutional	atom counts bond counts number of heavy atoms number of specific atoms ...	0D 0D 0D 0D	
Sub structural search	functional groups number of rings fragment counts fingerprints similarity diversity ...	2D 2D 2D/3D 2D/3D 2D/3D	Willet 2004
Graph Theory	molecular connectivity electrotological state indices (E-state) shape indices Wiener indices Balaban indices ...	2D 2D 2D 2D 2D	Kier and Hall 1999; Balaban 1997

(continued)

Table 1 (continued)

Information content/ Theory used	Representative descriptors	Dimensionality	Representative publications
Information theory	Entropy Shannon indices		Godden and Bajorath 2000
Autocorrelation descriptors 3D-Autocorrelation descriptors	Topological maximum cross correlation (TMACC) descriptors Maximum Auto-Cross-Correlation MACC descriptors	2D 3D	Spowage et al. 2009; Pastor et al. 2000
Structural keys	MACCS (MDL)	2D	MACCS 2011
QuaSAR descriptors	MOE descriptors (Molecular Operating Environment)	2D/3D	MOE-Chemical Computing Group
Statistical indices	WHIM descriptors GETAWAY descriptors	3D 3D	Gramatica 2006; Consomni et al. 2002

*Interaction Energy Fields value at each point of a Grid for each probe

**The information present in 3D maps, compressed into a few 2D numerical descriptors

Thus they provide insight into the mechanism of action. More to the point, easily calculated physicochemical and molecular properties have created the basis for the development of the drug-like concept (see Sect. 4.1.1).

On the other hand, theoretical descriptors may be considered to reflect a direct detailed representation of molecular structure. However they are not easily interpretable and they do not provide a straightforward perception of the mechanism of action. Their use in QSAR/QSPR models is often faced with some skepticism and their contribution to model quality and validity performance compared to classical descriptors has been questioned in the case of lead optimization (Vallianatou et al. 2013). However, it is true that in some cases the most predictive model may not be the most interpretable (Birchall et al. 2008a, b; Nicolotti et al. 2002). The value of models with high prediction accuracy but low interpretability has already been discussed in Sect. 3.

To obtain information about molecular structure from QSAR/QSPR models with low interpretability, a procedure called *reversible decoding* or *inverse QSAR* is being developed. Topological and molecular signature descriptors are considered to be more suitable for inverse QSAR/QSPR (Faulon et al. 2005; Gozalbes et al. 2002).

Moreover, sub-structural descriptors and molecular fingerprints are important to establish similarity/diversity approaches, which gain increasing interest within the scientific community (Willett 2004). Such approaches are widely used for virtual screening and design of chemical libraries, which aid in the primary identification of promising hits.

Recently, chemical similarity between molecules is being extended to evaluate clinical effects, if combined with information derived from computing similarity based upon lexical analysis of patient package inserts. It is expected, that drugs with highly structural similarity (both by 2D and 3D comparison) are much more likely to have significant overlap of their clinical effects, compared to drugs that are structurally different (low 2D similarity but high 3D similarity Yera et al. 2014). However in the search of new candidates chemical similarity does not always lead to biological similarity. Structure-Activity landscape may present the so called activity cliffs. Such discontinuities cannot be predicted by statistically derived QSAR models (Guha 2011).

In the case of toxicity predictions the incorporation of biodescriptors (short-term assays) as independent variables is suggested. Such descriptors are derived by in vitro quantitative high throughput screening (qHTS) and in combination with chemical descriptors lead to hybrid models, which may exhibit higher accuracy (Sedykh et al. 2011).

Gene expression signatures of a desired biological state, derived from gene expression data are used to screen a compound library to identify compounds that induce this target signature and corresponding phenotype, while they may also be used as descriptors (Hieronymus et al. 2006; Stegmaier et al. 2004).

4.1.1 Drug-Like Filtering

The use of combinatorial methods during the last 30 years has produced a vast number of compounds, which tend to be more lipophilic, less soluble and with higher molecular weight than conventional drug entities (Hertzberg and Pope 2000). Such properties are often associated with unfavorable absorption, poor or inconsistent bioavailability, as well as with lack of selectivity and increased toxicity (Oprea 2000). To face this situation the concept of druglikeness was launched, defining boundaries on the chemical space and functioning as filter to guarantee a physicochemical profile enabling further development (Leeson and Springthorpe 2007; Yusof et al. 2013). Druglikeness provides useful guidelines for early stage drug discovery, following simple rules of thumb, which suggest cut-off values or ranges for certain properties. According to the rule of 5 (RoF), molecular weight (MW) should not exceed 500 Da, calculated lipophilicity (clogP) should not exceed 5, hydrogen bond donor sites (HBD) should not be more than 5, and hydrogen bond acceptor (HBA) sites not more than 10. Upon pairwise violation of these limits, bioavailability problems may occur in the case of orally administered drugs (Lipinski et al. 1997). RoF was further extended including cutoff values or ranges for additional properties, the most common being: Polar Surface Area (PSA) < 140, number of rotatable bonds (ROTB) < 10, Molar Refractivity (MR) in the range of 40–130, number of aromatic rings (AROM) < 3, total number of atoms in the range of 20–70 (Ursu et al. 2011; Veber et al. 2002). Lipophilicity is related also to safety endpoints. Increased relative risk (6:1) for an adverse event may be anticipated for compounds possessing high lipophilicity (ClogP > 3) and low topological polar surface area (TPSA < 75 Å) (Hughes et al. 2008). It is also reported that for ClogP > 3 there is a dramatic higher risk for hERG channel inhibition, an endpoint associated with cardiotoxicity (Wager et al. 2011). More strict cutoff values are proposed for compounds intended to act in the Central Nervous System (CNS-likeness Pajouhesh and Lenz 2005). A quantitative estimate of drug-likeness (QED) has been proposed by Bickerton et al. (Bickerton et al. 2012) which relates the similarity of a compound's properties to those of oral drugs based on eight commonly used molecular properties: MW, log P, HBDs, HBAs, PSA, ROTBs, AROMs and count of alerts for undesirable substructures.

For lead compounds the rule of 3 is suggested according to which MW < 300, logP < 3, HD < 3, and HA < 6 (Congreve et al. 2003). The rule of 3 is applicable mainly for fragment-based lead generation.

The rules of thumb are very simple and understandable, however they do not take into account inaccuracies in the prediction of logP and more important they do not consider the receptor demands. For instance, receptors of the PPAR family possess a very large hydrophobic cavity in their active center, requiring lipophilic ligands with high molecular weight, which in many cases violate twice the rule of 5 (Giaginis et al. 2008, 2007). Target specific lipophilicity profiles obtained through calculation of the logP and logD of ligand series for different receptors have recently investigated, showing also other targets where the compound libraries had mean logP \geq 5, i.e., outside of traditional RoF space with respect to lipophilicity.

Such knowledge in the early stages of drug development is very useful for the formulation strategy in later stage (Bergström et al. 2016).

The advantages of smaller and less lipophilic compounds as safer and more selective drug candidates were further recognized in terms of receptor binding. According to metrics such as ligand efficiency (LE) and ligand lipophilicity efficiency (LLE) affinity is normalized against molecular size, expressed as heavy atoms, or lipophilicity respectively (Abad-Zapatero 2007; Hopkins et al. 2014). Ligand efficiency dependent lipophilicity (LELP) takes both lipophilicity and molecular size into consideration by dividing logP (clogP) by LE (Tarcsay et al. 2012). In terms of thermodynamics, according to the above metrics drug—receptor binding should be optimized in regard rather to the enthalpic component through specific interactions. Such metrics may be used to prioritize drug candidates with quasi equal potency (Hann 2011; Leeson and Springthorpe 2007).

An update on recent applications of efficiency metrics and strategies to control drug-like properties and to replace problematic elements for improving drug design, is recently published by Meanwell (2016).

4.2 Modeling Techniques

Statistical tools mediate the relationship between structural descriptors and the response variable(s) leading either to regression or to classification models. Model building methods are incorporated in different software packages (Bruce et al. 2007). Multiple linear regression (MLR) analysis is a simple and still widely used technique, which however can handle a limited number of variables. Thus, as a first step, variable selection methods are applied to reduce the large number of calculated descriptors to a set which is information rich but as small as possible. Redundant descriptors and descriptors which show low variance or/and collinearity are removed. For further descriptor reduction, stepwise regression approaches are commonly used, with the drawback however that they are local search processes and may converge to local optima (Paterlini and Minerva 2010).

A promising alternative for variable selection is the use of genetic algorithms (GA). GAs explore the descriptor space simultaneously by a population of candidate solutions which compete and recombine, mimicking the process of natural selection (Mitchell 1998).

Reduction of the descriptors space is inherent in multivariate data analysis (MDVA) a popular statistical technique, which permits the simultaneous (not one at a time) treatment of large number of descriptors, while tolerating inter-relation between them (Eriksson et al. 2001; Wold et al. 2001). It is a projection method from a space with high dimensionality to a space with few dimensions (latent variables), characterized as principal components. Principal component analysis (PCA) is a powerful unsupervised classification method. Projection to latent structures defined also as partial least squares (PLS) is the regression extension of PCA. PLS can handle more than one response variables, under the precondition that

they are to some degree inter-related. This is very important for multi-target drug design, for toxicity models or for the establishment of activity profiles of antimicrobial or anticancer agents (Vallianatou et al. 2013; Koukoulitsa et al. 2009). PLS analysis generates coefficients for the original variables (descriptors), which permit a straight-forward interpretation of the model.

MLR and PLS are linear methods and any non-linearity should be incorporated through data transformation before the analysis. On the other hand, machine learning (ML) methods are gaining increasingly important roles in the construction of classification and/or prediction models in several steps of the drug discovery process (Tao et al. 2015). They are effective dimension reduction methods, while allowing for non-linearity to be included in the models and the incorporation of variable interactions. Thus they can reflect biological complexity leading to models with high accuracy. Their drawback is their black box character, e.g., the inability for their rationalization and interpretation in chemical terms. Most popular ML techniques are artificial neural networks (ANN) and associative neural networks (ASNN), inspired by the function and structure of neural network correlations in brain, the k-nearest neighbor technique (k-NN), support vector machines (SVM), regression trees (RT) or random forest (RF) (Byvatov et al. 2003; Sakiyama 2009). The latter are also very useful in the creation of gene expression signatures (Lima et al. 2016). An overview of the machine learning methods, used mainly as prediction tools for ADME properties is given in a recent review by Tao et al. (Tao et al. 2015). Table 2 includes commonly used statistical tools, which are referred in the representative QSAR and QSPR examples, discussed in Sect. 5.

Models are evaluated by statistical data, the most commonly being correlation coefficient (R or r) and determination coefficient (R^2 or r^2), standard error of estimate(s), given also as root mean square error of estimate (RMSE). The adjusted determination coefficient (R_{adj}^2) for degrees of freedom allows for comparison between QSARs with different numbers of descriptors and can indicate if a given QSAR model is overfit incorporating too many descriptors. The Fisher test F provides an indication of a chance correlation, while the Student test t is used to evaluate the significance of descriptors in MLR. In multivariate data analysis, the variable importance to projection (VIP) criterion is used instead. In ANN, the contribution of molecular descriptors is based on the ratio between the performance of neural network before and after the elimination of each descriptor (sensitivity analysis).

Visualization of the results, fitting the line on the graph of observed versus predicted values, enables to check for outliers or trends in the data, while it provides an overview of the predictive power of the model. In fact a good model should show an 1:1 correlation between observed and predicted values. Detected outliers should be submitted to further investigation—they may unravel interesting information. Further statistical data are related to model internal or external validation (Sect. 4.3).

For classification models, % sensitivity defined as the ratio of percentage of true positives in respect to the sum of true positives + false negatives, % specificity, defined as the ratio of percentage of true negatives in respect to the sum of true

Table 2 Statistical tools, commonly used in QSAR/QSPR prediction or classification models

	Linear	Non-linear	Prediction	Classification
Multiple Linear Regression Analysis (MLR)	x		x	
Partial Least Square/Projection latent Structures, PLS	x		x	
Principal Component Regression, PCR	x		x	
Principal Component Analysis, PCA	x			x(unsupervised)
PLS-Discriminant Analysis, PLS-DA	x			x (supervised)
Linear Discriminant Analysis (LDA)	x			x (supervised)
Artificial neural networks, ANN Bayesian NN Associative NN		x	x	x (unsupervised/ supervised)
Support Vector Machine, SVM		x	x	x (supervised)
k-Nearest Neighbors		non-parametric	x	x (supervised)
K-means		x	x	x(unsupervised)
Decision trees and Random forests		x	x	x(unsupervised)
Classification and Regression Tree (CART)		x	x	x (supervised)
Ensemble methods-Bagging- Boosting trees		x	x	x (supervised)

negatives + false positives and %CCR (correct classification rate or balanced accuracy) equal to (sensitivity + specificity)/2 are common statistical data to evaluate the merit of the models. It should be noted that acceptance criteria depend on the quality of experimental data, as well as on the ultimate goal of the QSAR/QSPR performed.

4.3 Model Validation

Whatever modeling technique is used, validation of QSAR models has received considerable attention in the last decades (Guha and Jurs 2005; Tropsha et al. 2003; Veerasamy et al. 2011). Validation requirements are becoming increasingly strict so as to assure robust models, which can lead to reliable predictions and to proof of concepts. According to the European center for the validation of alternative methods (ECVAM) four tools, the methods accepted for estimating the prediction accuracy include (i) cross-validation, (ii) bootstrapping, (iii) randomization of the response data, and (iv) external validation (Worth et al. 2004).

Cross-validation as an internal model validation method is usually performed by the 'leave-one out' (LOO) or 'leave many out' (LMO) procedure to determine PRESS and cross-validated correlation coefficient q^2 , which are metrics reflecting

the internal predictive ability of the model. In contrast to r^2 which increases with the number of variables included in the model with a tendency to approximate the value of 1, Q^2 follows a quadratic relationship reaching a maximum corresponding to optimal number of variables.

To check that the obtained model is not a result of chance factors, randomization of the Y response is recommended (Rücker et al. 2007). All models obtained with the randomized training set should be inferior, with r^2 and q^2 values around 0 or with negative values respectively for a set with 0% similarity with the original set (Gasteiger et al. 2003; Klopman and Kalos 1985).

A prerequisite for model validation is external validation, either by dividing the data set into training and test sets and rebuilding the models or/and using a blind test set. The errors produced in the predictions should be comparable to those achieved for the training set. Recently, Roy et al. have proposed a modified correlation coefficient r_m^2 as a novel metric for external validation, which represents the actual difference between the observed and predicted response data without consideration of training set mean and taking into account the r^2 with intercept and r_0^2 , without intercept. Change of the axes denoting observed and predicted y modified correlation coefficient $r_m'^2$ may be different from r_m^2 . A threshold for the difference $\Delta r_m^2 = \text{abs}(r_m^2 - r_m'^2)$ less than 0.2 and an average $r_m^2 = (r_m^2 + r_m'^2)/2$ higher than 0.5 indicate robustness of the model (Roy et al. 2009; Roy et al. 2012).

Model applicability domain (AD), defined as the region of chemical space where predictions can be made without extrapolation is an important issue that should be taken into consideration for the proper use of QSAR/QSPR. There are different methods for the assessment of applicability domain, for particular types of QSAR models (Jaworska et al. 2005; Netzeva et al. 2005; Sahigara et al. 2012). Distance/leverage based methods are usually applied. In regard to QSAR models for regulatory purposes, OECD clearly states that the AD should be described “*in terms of the most relevant parameters, i.e., usually those that are descriptors of the model*” (Jaworska et al. 2003).

The performance of the models over time, in particular in the case of global QSPR models, has been addressed by continuous updating of the original models, so as to extend the applicability domain allowing predictions for new compounds of different chemotypes (Rodgers et al. 2011).

5 QSAR/QSPR Applications in the Drug Discovery Process

QSAR/QSPR models can be established for all processes across the drug discovery pipeline. Initial virtual screening may be followed by modeling of the affinity of ligand series to the receptor or to other off-target proteins. In parallel, models for permeability and other pharmacokinetic properties like plasma protein binding, affinity to uptake or efflux transporters and metabolic stability may be established to evaluate safety and efficacy of the candidates.

5.1 Modeling Pharmacodynamics

Pharmacodynamic models focus on predictions of receptor affinity. It should be noted however that binding to proteins is governed by the same recognition forces, regardless if they are target receptors, plasma and tissue proteins, metabolizing enzymes or off-target proteins. They reflect interactions between the small molecules and the amino acid residues within the active site of the protein.

Computational techniques to detect and/or optimize efficient binding involve both ligand- and structure-based methods and are applied to optimize receptor binding as well as to predict ADME properties involving proteins, like plasma protein binding, binding to metabolizing enzymes or transporters (Fig. 2).

5.1.1 Ligand-Based Drug Design (LBDD)

Ligand-based Quantitative Structure-Activity Relationships (QSAR), established by the procedures, already discussed in Sects. 3–5, do not require or ignore knowledge on the structure of the target protein. In most cases, they are two dimensional models, although they may embrace three dimensional information by incorporating 3-D descriptors. Such models take advantage of the large number of available descriptors and the progress in the statistical techniques as well as of the associated philosophy (see Sect. 4). They can be further classified as global or local models.

Global models are useful for virtual screening, off target screening or for plasma/tissue protein binding (Helgee et al. 2010; Sheridan 2014). For global models, the goal is to encompass a large applicability domain, while interpretability may not be an issue, at least in the early stages. More important may be the continuous updating of the models to incorporate new chemotypes, so as to expand their applicability domain (Rodgers et al. 2011). In fact, the goal of such global models is not the search for new chemical entities, but to prioritize existing or virtual compounds. In contrast, for lead optimization on receptor binding, local models are more helpful. They are built under the precondition that all analyzed molecules interact with the same type of receptor in the same manner. Evidently, in these cases interpretability defines a determinant factor since the primary goal is to understand the receptor requirements and search for novel compounds with the desired physicochemical/molecular properties. Yet, the inverse-QSAR methodology (see Sect. 4), although based on descriptors which do not confer interpretability, may still allow to construct viable molecules (Wong and Burkowski 2009).

The three dimensional structure of the molecules can serve to create 3-D QSAR models, which provide a direct link to potency. 3D-QSAR has emerged as an extension to the classical 2D-QSAR, using robust chemometric techniques, such as PLS. In 3D-QSAR the precondition for identical binding sites in the same relative

geometry for all molecules should be strictly obeyed. After geometry optimization, molecules are superimposed and carefully, aligned in a rational and consistent way to create a hypermolecule. A sufficiently large box is positioned around this hypermolecule and a grid distance is defined. Different atomic probes, e.g., a carbon atom, a positively or negatively charged atom, a hydrogen bond donor or acceptor, or a lipophilic probe, are used to calculate field values in each grid point, i.e., the energy values which the probe would experience in the corresponding position of the regular 3D lattice. Using these fields as input descriptors in PLS analysis, principal components, defined by different proportions of the fields, are generated.

The most popular 3D-QSAR methodology is Comparative Molecular Field Analysis (CoMFA) and Comparative Molecular Similarity Indices Analysis (CoMSIA). CoMFA, developed by Cramer in 1988, is based upon the calculated energies of steric and electrostatic fields (Cramer et al. 1988). CoMSIA, instead of interaction fields, calculates similarity indices using a distance-depended Gaussian functional form. Five types of similarity indices are calculated, steric, electrostatic, hydrophobic, and hydrogen-bond donor and acceptor (Klebe 1998). An important advantage of CoMFA and CoMSIA is the graphical representation of the results. 3-D contour maps in CoMFA display the different contributions of the potentials to the activity, while in CoMSIA they highlight the areas within the region occupied by the ligands, that 'favor' or 'dislike' the presence of a structural feature with a given physicochemical property. In this sense the CoMSIA representation is more easily interpretable than CoMFA contour maps.

The difficulties of both methods are associated with the structure alignment, which may affect the results, while it limits their application to strictly similar compounds. The use of a single conformation for a given ligand represents a limitation of 3D-QSAR since the bioactive conformation may not be necessarily the thermodynamically optimal one. Moreover, orientation in the binding site may be ambiguous, especially in the absence of structural information on the biological receptor. To face such problems, higher dimension QSAR methodologies (xD-QSAR) have been developed. Additional dimensions offer the possibility to represent each ligand molecule as an ensemble of conformations, orientations, tautomeric forms and protonation states (Ekins et al. 1999; Hopfinger et al. 1997; Vedani et al. 2005, 2000).

A general drawback of ligand-based QSAR models is the underlying assumption that chemical similarity correlates with biological similarity, considering a rather smooth structure-activity landscape. The presence of activity outliers however shows that this is not always the case and structure-activity landscape may present activity cliffs (Guha 2011). In such cases, outliers deserve special attention and should be investigated separately. Outliers representing activity cliffs can be identified by structure-based methods, like docking or pharmacophore approaches. In this aspect combination of both ligand- and structure-based approaches may provide insight on the behavior of such outliers (Vallianatou et al. 2013).

5.1.2 Structure-Based Drug Design (SBDD)

Structure-based methods rely on detailed knowledge of target protein structures and target protein-ligand complex providing a more straightforward understanding of the mechanistic aspects in drug-receptor interactions. X-ray crystallography as well as NMR have contributed immensely in this field (Anderson 2003).

In the PDB database (<http://rcsb.org>), more than 120,000 biological macromolecular structures are deposited, covering more than 40,000 organisms and 38,000 distinct protein sequences. However, in order to use those data, a proper and detailed preparation of the protein must be performed (Anderson 2003; Sastry et al. 2013). The preparation process includes hydrogen addition, protonation or deprotonation based on pK_a prediction of acid or basic side chains, and side chain optimization to achieve the optimum number of hydrogen bond interactions. Once the structure of the protein is well studied and analyzed, all essential parts for interactions between the co-crystallized ligand and the protein are gathered to design new optimized molecules. In this aspect, the key issue for a successful structure-based design is the identification of the target and the appropriate binding site. In Fig. 3 a representative crystal structure of a protein-ligand complex and the interaction points is illustrated. In Fig. 3a, PPAR α receptor is represented by ribbons in complex with aleglitazar, represented in space-filling way (CPK representation). Figure 3b shows the ligand interaction diagram of aleglitazar inside the binding pocket.

Additionally, the crystal structure of a protein-target can be used for virtual screening procedure. Virtual screening procedures are based on the structure of a protein while a large database is screened and all molecules are ranked based on empirical docking scoring function for binding affinity (Hillisch et al. 2015). Top ranked molecules are then tested in vitro to validate the model, and the new lead compounds are optimized using computer-aided combinatorial techniques (CombiGlide, version 4.1, Schrödinger, LLC, New York, NY, 2016). Thus, using fragment based algorithms, new virtual chemical libraries are designed based on the core skeleton of the hit compound previous, and top ranked “theoretical” molecules are passed to medicinal chemists for synthesis and further in vitro testing.

However prediction of binding constants based on the correlation with docking scores is not always feasible, especially in the case of structurally diverse compounds. ΔG values calculated by molecular docking may have an acceptable calculation error of 2 kcal/mol corresponding to 2 log units of dissociation constants K_d (Enyedy and Egan 2008; Keserü 2001). Moreover, they may show little differentiation, since they are the outcome of enthalpy–entropy compensation (Brandt et al. 2011). Therefore docking calculations alone are not sufficient, if the principal query is to predict binding constants.

In the past years, many success stories have been achieved using structure-based drug design (SBDD). Some representative examples are reported below:

Amprenavir (Agenerae) and nelfinavir (Viracept) (Kaldor et al. 1997) were the first drugs reaching the market designed against HIV protease using SBDD methodology. Zanamivir (Relenza) was designed against neuraminidase (Varghese 1999),

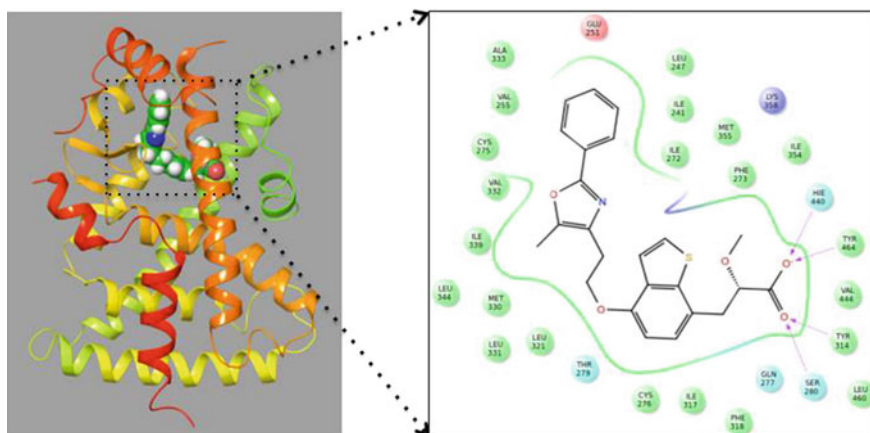


Fig. 3 **a** Ribbon representation of PPAR α in complex with aleglitazar (CPK representation), **b** Ligand interaction diagram of Aleglitazar inside the binding pocket. Hydrophobic residues are colored *green*, hydrophilic residues are colored *cyan*, positive charged residues are colored *blue* and negative charged residues are colored *red*. Hydrogen bonds are depicted with *dashed lines*

Tomudex against thymidylate synthase (Rutenber and Stroud 1996) and imitin-abmesylate (Glivec) against Abl tyrosine kinase (Schindler et al. 2000). Moreover, SBDD has contributed to address more complicated targets, like nucleic acids as well as protein-protein interactions. Thus, inhibitors have been developed for HIV-1 RNA target TAR (Lind et al. 2002, Filikov et al. 2000), the IL2/IL2R α receptor interaction (Tilley et al. 1997), the VEGF/VEGF receptor (Wiesmann et al. 1998) and Bcl2 (Enyedy et al. 2001).

5.2 Modeling Pharmacokinetics

Pharmacokinetic processes are controlled both by passive phenomena and binding to proteins, the latter concerning plasma and tissue proteins, metabolizing enzymes and transporters. Passive phenomena include passive diffusion through various biological barriers, hemolysis or cell retention. They are governed primarily by lipophilicity, while molecular weight and hydrogen bonding may contribute as additional factors (Avdeef 2012; van de Waterbeemd and Smith 2001). There are also border cases between passive diffusion and binding such as phospholipidosis or drug membrane interactions (Hanumegowda et al. 2010). Volume of distribution is also the outcome of membrane permeability and tissue binding (Hollósy et al. 2006). Among the biological barriers, the gastrointestinal tract and the blood brain barrier are of highest interest and relevant QSPR models are discussed in the following sections.

5.2.1 Modeling Permeability

Several in vitro techniques have been developed for rapid estimation of membrane permeability in vitro. Artificial membranes used in parallel artificial membrane permeability assay (PAMPA) (Kansy et al. 1998) or in immobilized artificial membrane (IAM) chromatography (Tsopelas et al. 2016a, b) provide easy measurements. However, cell-based protocols such as Caco2 or MDCK cell lines are more widely accepted as measures of effective permeability, which is considered as a reliable index mainly for intestinal human absorption (Thiel-Demby et al. 2008; Usansky and Sinko 2005; Volpe 2008; Yee 1997). The Caco-2 model is recommended by the US FDA for the classification of compounds according to the bio-classification system (BCS) (Larregieu and Benet 2013). Several QSPR models to predict Caco-2 or MDCK permeability have been published, which however include a limited number of compounds (Castillo-Garit et al. 2008; Irvine et al. 1999; van De Waterbeemd et al. 1996). It has been shown however from local models, that high Caco-2 permeability rate should correspond to the high human intestinal permeability rate (or extent of absorption), independent of the laboratories of origin and regardless of whether carrier-mediated transport is occurring (Larregieu and Benet 2014).

Due to the considerable inter- and intra-laboratory variability of Caco-2 effective permeability, classification models may be a better option, while meeting the requirements for BCS. Two representative studies performed on large datasets are reported below. Sherrer et al. applied random forest (RF) to the largest dataset ever reported (15791 compounds) to establish a moderate model with a $R^2 = 0.52$, $RMSE = 0.20$ using 8 descriptors (Sherrer et al. 2012). A later model derived by ruled-based decision trees using 1289 compounds achieved determination of 3 permeability classes (High-H, Medium-M, Low-L). The best rule, based on the combination of $PSA-MW-\log D$ (3P Rule), was able to identify the H, M and L classes with accuracy of 72.2, 72.9 and 70.6%, respectively, while a consensus system based on three voting binary classification trees predicted 78.4/76.1/79.1% of H/M/L compounds on the training and 78.6/71.1/77.6% on the test set (Pham-The et al. 2013).

Recently, a QSPR study to predict Caco-2 cell permeability was performed on a large data set of 1272 compounds, which were filtered and curated (Wang et al. 2016). Four different methods including multiple linear regression (MLR), partial least squares (PLS), support vector machine (SVM) regression, and boosting trees were employed to build prediction models with 30 molecular descriptors. The nonlinear model derived by Boosting performed better with $R^2 = 0.97$, $RMSE = 0.12$, $Q^2 = 0.83$, $RMSECV = 0.31$ for the training set and $R^2 = 0.81$, $RMSE = 0.31$ for the test set.

5.2.2 Predicting Human Intestinal Absorption/Oral Bioavailability

Considerable efforts are oriented to establish QSPR models for human intestinal absorption and oral bioavailability. Relevant software packages are available either for direct predictions or for predictions of ADME properties like lipophilicity, solubility, ionization, which would allow a rough evaluation of the potential of drugs to be orally absorbed. The rules of thumb, discussed in Sect. 4.1, are very helpful in this case.

Human intestinal absorption (HIA) is usually measured as the percentage of the dose that reaches the portal vein after passing the intestinal wall (%HIA). On the other hand, oral bioavailability (%F) describes the passage of a substance from the site of absorption into the systemic circulation after first pass hepatic metabolism. Intestinal metabolism, acidic stability and the effect of transporters contribute to the outcome. Absorption in gastrointestinal tract is governed by permeability through cell membranes (transcellular absorption) or through the intercellular space between cells of the gastrointestinal mucosa (paracellular transport). The effect of lipophilicity on absorption has been previously described by linear, bilinear, sigmoidal or parabolic models (Kubinyi et al. 1993; Kubinyi and Kehrhan 1978). However, for the establishment of global QSPR models, which would permit predictions for different chemotypes of novel compounds, additional physico-chemical parameters or molecular descriptors, should be implemented. Molecular weight, polarity or hydrogen bonding parameters as well as the charge state are most commonly used, being also consistent to describe Caco-2 permeability as discussed above (Kumar et al. 2011; Tsopelas et al. 2016a, b; Veber et al. 2002).

The main problems to be addressed for the establishment of robust global HIA models concern the significant variability of the datasets from one source to another and the distribution of endpoints, since they include commercially available drugs and are often heavily biased towards compounds with high intestinal absorption values (Hou et al. 2007). This fact will influence the predictive capacity of the *in silico* models and better predictions will be obtained for compounds with high intestinal absorption values, compared to the rest of the dataset. A scientific and technical report of the European Commission Joint Research Centre and the Institute for Health and Consumer Protection compiles literature models for HIA published till 2010, along with databases with ADME endpoints (Mostrag-Szlichtyng and Worth 2010). In this chapter, representative examples and latest investigations are discussed.

One of the first attempts to predict %HIA was published by Wessel et al. who applied a genetic algorithm with a neural network (GA-NN) technique to develop a non-linear model for set of 86 drugs. They identified six most significant variables, namely: the cube root of gravitational index, related to the size of molecule, the normalized 2D projection of the molecule on the YZ plane (SHDW-6, related to the shape, the number of single bonds (NSB), related to flexibility, as well as the charge on hydrogen bond donor atoms (CHDH-1), the surface area multiplied by the charge of hydrogen bond acceptor atoms (SCAA-s) and the surface area of hydrogen bond acceptor atoms (SAAA-2), related to hydrogen-bonding properties.

The predicted %HIA values achieved good statistics with root mean square errors (RMSE) of 9.4%HIA units for the training set, 19.7%HIA units for the cross-validation (CV) set, and 16.0%HIA units for the external prediction set (Wessel et al. 1998).

The general solvation equation developed by Abraham's group (Abraham et al. 2002) was used by Zhao et al. to model the human intestinal absorption data of 169 drugs (Zhao et al. 2001). The model Eq. (3) derived by stepwise MLR was based on Abraham's linear solvation energy (LSE) descriptors, namely: excess molar refraction (E), solute polarity/polarizability (S), the McGowan characteristic volume (V), solute overall hydrogen bond acidity (A) and basicity (B).

$$\begin{aligned} \%HIA &= 92 + 2.94E + 4.10S + 10.6V - 21.7A - 21.1B \\ R^2 &= 0.74, s = 14 \end{aligned} \quad (3)$$

According to Eq. (3) the volume and the hydrogen bond descriptors were found to be the most important.

Klopman et al. compiled a large dataset of 467 drug molecules for human intestinal absorption. The data were split into a training set of 417 and external prediction set of 50 molecules. Structural fragments promoting or preventing HIA were identified using the CASE program (<http://www.multicase.com/>) and their occurrence was subsequently used in a multiparameter linear equation (4) to predict human intestinal absorption (Klopman et al. 2002).

$$\%HIA = c_0 + c_i G_i, \quad (4)$$

where c_0 is a constant, c_i are the regression coefficients and G_i is the presence (1) or absence (0) of a certain structural fragment. The final QSAR model included 37 descriptors: 36 statistically significant structural descriptors identified by CASE analysis and one important physicochemical parameter—the number of hydrogen bond donors (H donors). The model was able to predict the %HIA with an $r^2 = 0.79$ and a standard deviation $s = 12.32\%$ for the compounds of the training set. The standard deviation for the external test set (50 drugs) was 12.34%. The merit of the model is that it indicates certain substructures with negative impact in %HIA, such as quaternary nitrogens, SO_2 groups connected to an aromatic ring and others with positive impact on HIA. A drawback of the model is that the training set was biased towards high absorption values (Klopman et al. 2002).

Using Zhao's data set, Sun proposed a PLS-DA classification approach for human intestinal absorption modeling, using atom type descriptors. Drugs were classified as classified them as "good" (absorption > 80%) "medium" (80% < absorption < 20%) or "poor" (absorption < 20%), according to their %HIA. A five component PLS-DA model separated very well all 169 compounds with $r^2 = 0.921$ and $q^2 = 0.787$. Since in the case of virtual screening, only poorly absorbed compounds would need to be identified and removed the authors proposed also a

three-component PLS-DA with $r^2 = 0.939$ and $q^2 = 0.861$ to separate the compounds with less than 20% absorption (Sun 2004).

Recently, a dataset of 578 compounds, split into a training set of 403 compounds a validation set of 87 and an external prediction set of 87, was analyzed, using ensemble learning (EL) techniques, (gradient boosted tree, GBT and bagged decision tree, BDT) to derive both qualitative (classification) and quantitative models. Topological polar surface area proved to be the most important descriptor with negative contribution, followed by lipophilicity expressed as XlogP. Classification accuracy $> 99\%$ was reported, while the QSAR models yielded correlation coefficients $R^2 > 0.91$ between the measured and predicted HIA values (Basant et al. 2016).

Prediction models are available also for the more complex process of oral bioavailability (Andrews et al. 2000; Hou et al. 2007; Kim et al. 2014; Kumar et al. 2011; Martin 2005; Moda et al. 2007; Tian et al. 2011). Till the year 2010 they are compiled in the scientific and technical report of the Joint Research Center of the European Union. In the same report relevant software for prediction of oral bioavailability are provided (Mostrag-Szlichtyng and Worth 2010).

Recently, *in silico* approaches focus more on physiologically based pharmacokinetics (PBPK), which go beyond human intestinal absorption and oral bioavailability, providing realistic descriptions of absorption, distribution, metabolism, and excretion processes (Bois and Brochot 2016; Jamei 2016). PBPK modeling has gained a significant impact on regulatory science and decisions (Huang et al. 2013) and best practice for its use to address regulatory questions, has been reported (Zhao et al. 2012).

5.2.3 Predicting Blood Brain Barrier Penetration

In drug discovery for CNS active drugs, it is important to determine whether a candidate molecule is capable of penetrating the blood brain barrier (BBB). For drugs targeted at the CNS, the BBB penetration is a necessity, whereas for drugs acting in peripheral tissues, the BBB penetration may lead to undesirable adverse effects (Di et al. 2009; Ecker and Noe 2004). The log BB, defined as the logarithm of the ratio of the concentration of a drug in the brain and in the blood, measured at equilibrium, is an index of BBB permeability. The optimal threshold for classification as a CNS acting drug is typically specified between 0 and -1 (Clark 2003). Log BB values, although widely used, do not take into account plasma and tissue binding, and therefore, do not reflect the free amount of the drug in the brain. Permeability surface area product (PS, quantified as logPS) representing the uptake clearance across the BB is used as a direct measure of permeability and theoretically is not confounded by the plasma and brain tissue binding.

Several models have been published trying to predict blood-brain barrier permeability from various physicochemical properties of molecules, including, among others, molecular size, lipophilicity or number of groups that can establish potential hydrogen bonds (Clark 1999; Kaliszan and Markuszewski 1996;

Konovalov et al. 2007; Luco 1999; Vastag and Keseru 2009). Rules of thumbs are also suggested, as discussed in Sect. 4.1. Till the year 2010, literature models are compiled in the scientific and technical report of the European Commission Joint Research Centre and the Institute for Health and Consumer Protection (Mostrag-Szlichtyng and Worth 2010). Some representative models and recent publications are discussed in this chapter.

Already in 1980, Levin had related $\log P_c$ (which is close analog of $\log PS$) to a simple linear function of $\log P$ and molecular weight. The overall effect was represented as $\log(P * MW^{-1/2}) = \log P - 1/2 \log MW$, whereby increasing $\log P$ was supposed to reflect a steady increasing $\log PS$ effect, whereas increasing MW had an opposite effect (Levin 1980). In 1999, Clark analyzed a set of 55 diverse organic compounds and generated a multiple linear regression model based on *in silico* calculated polar surface area (PSA) and $\log P$ values with negative and positive contribution respectively (Clark 1999).

The linear solvation energy relationship approach (LSER), also used to model human intestinal absorption, has been applied to blood/brain permeability prediction (Platts et al. 2001). For a dataset of 148 diverse compounds using MLR, they obtained a transparent QSAR incorporating 5 Abraham descriptors and an indicator variable (equal 1 for carboxylic acids and 0 for other compounds) has been reported. The model shows good statistics ($R^2 = 0.74$, $s = 0.34$, $RCV^2 = 0.71$). According to the model, the increasing size of molecules strongly enhances brain uptake, while increasing polarity/polarizability, hydrogen-bond acidity, basicity and the presence of carboxylic acid groups have a detrimental effect. Platt's model has been implemented in the commercially available ADME Boxes software (previously Pharma Algorithms; now ACD Labs, <http://www.acdlabs.com/>), providing a very fast estimation of $\log BB$. Later, the data set was extended to include 328 compounds with *in vivo* and *in vitro* $\log BB$ values. A correlation coefficient $r^2 = 0.75$ and a standard deviation $s = 0.3$ was achieved by incorporating an additional indicator for *in vitro* data (Abraham et al. 2006).

For a data set of 88 diverse compounds using a variable selection and modeling method, a QSAR with three or four descriptors out of 324 descriptors has been reported for $\log BB$ prediction. In both models, calculated lipophilicity (AlogP98) was combined either with the atomic type E-state index (SsssN) and Van der Waal's surface ($r = 0.842$, $q = 0.823$, and $s = 0.416$) or with kappa shape index of order 1, atomic type E-state index (SsssN), atomic level based AI topological descriptor (AISsssC) ($r = 0.864$, $q = 0.847$, and $SE = 0.392$). The success rate of the reported models in test sets was 82% in the case of BBB + compounds. A similar success rate was observed with BBB-compounds (Narayanan and Gunturi 2005).

The VolSurf technique, which is based on molecular interaction fields, has also been used for blood/brain partitioning modeling (Crivori et al. 2000). The model was built on the basis of 230 diverse compounds and more than 70 VolSurf descriptors. Its prediction accuracy (assessed against an external test set) is 90% for BBB permeable molecules and 60% for non-permeable ones. The computational procedure is fully automated and fast and it provides a valuable tool for the virtual

screening of large datasets of diverse molecules (Cruciani et al. 2000). The shortcoming of this approach however is its low interpretability.

Linear discriminant analysis (LDA) based on physicochemical descriptors calculated *in silico* has been used to establish two distinct classification models (Vilar et al. 2010). The data set consisted of the 307 compounds used by Abraham et al. (Abraham et al. 2006) for which *in vivo* logBB values were available. Considering that molecules with $\log BB > 0.3$ cross the BBB readily while molecules with $\log BB < -1$ are poorly distributed to the brain, these values were selected thresholds for classifying the compounds into two categories. For the threshold 0.3, a two component model was obtained with lipophilicity and topological polar surface area (TPSA), the latter with a negative coefficient. For the threshold-1, the total number of acidic and basic atoms was additionally incorporated, also with a negative sign. The models were validated with external data sets using the area under receiver operating characteristic (ROC) curves as evaluation criterion. In ROC the fraction of true positives (sensitivity) is plotted against the fraction of false positives (1-specificity). An area under the ROC curve of 0.95 for model 1 and 0.97 for model 2 is reported, demonstrating the high predictive power of the models, considering that for a perfect classifier the area under the curve is 1 and for a random classifier it is 0.5 (Vilar et al. 2010).

Based on logPS values in rats, Suenderhauf et al. developed predictive computational models (decision tree induction) for a dataset of 153 compounds. The established models exhibited a corrected classification rate of 90%. The models confirmed the involvement of lipophilicity, molecular size and charge in BBB permeation (Suenderhauf et al. 2012).

5.2.4 Modeling Plasma Protein Binding

A special case of binding of small molecules to macromolecules is plasma protein binding. Plasma protein binding (PPB) is the reversible association of a drug with the proteins of the plasma and is mainly due to hydrophobic and electrostatic interactions. Since only the fraction of unbound (*f_u*) drug is able to pass across cell membranes, PPB strongly influences volume of distribution, half-life and efficacy of drugs. Extended plasma protein binding may be associated with drug safety issues, low clearance, low brain penetration, as well as drug–drug interactions (Ito et al. 1998; Rowley et al. 1997). In fact, plasma protein binding belongs to the ADME properties, representing mainly the “D” of the acronym.

Among the plasma proteins, human serum albumin (HSA) has a central role and the affinity of drugs to this protein is considered to dominate PPB and the thereupon related pharmacokinetic issues. Two primary active sites on HSA have been recognized for drug binding, the Sudlow’s sites 1 (warfarin site) and 2 (benzodiazepine site), α_1 -acid glycoprotein (AGP) is the second essential plasma protein with two main variants and a complicated physiological role (Lambrinidis et al. 2015).

Modeling of total plasma protein binding or/and of HSA binding has been the objective of many researchers and offers a representative case where combined

structure- and ligand-based methods act synergistically. Structure based methods are very helpful to initially classify the compounds according to the preferred binding site or protein, prior to proceeding to ligand-based methods. Since PPB is practically involved in any class of therapeutics, the ultimate goal is to construct global HSA or PPB models, where structural diversity plays an important role. Representative successful efforts are described below. Often more than one model are suggested by the same research group, where interpretability may compete with accuracy in predictions.

A multiple computer-automated structure evaluation method (M-CASE) was used by Saiakhov et al. (Saiakhov et al. 2000) to analyze 154 structurally diverse compounds for total plasma protein binding. M-CASE starts by searching for 'baseline correlation' via an internal baseline activity identification algorithm subroutine (BAIA), using the octanol-water partition coefficient which is the most important parameter. For compounds showing residual binding when predicted by the baseline correlation, the algorithm continues to identify responsible structural characteristics, called biophores. Several local QSAR models built for subsets with common biophores are included in the final global model. The binding site(s) of each biophore, including the warfarin, benzodiazepine and digitoxin sites, as well as AGP and lipoproteins, are also characterized. Lipophilicity as the prevalent parameter showed different contribution in each local QSAR, indicating different lipophilicity requirements for each binding site. A crucial structural fragment present in the molecules was found to be part of a phenyl ring. The model, after classifying the compounds according to their biophores, was able to predict correctly the percentage bound to plasma for 80% of the compounds with an average error of 14%.

A large data set of 1008 compounds, partitioned into a training set of 808 compounds and an external validation test set of 200 compounds was used by Votano et al. for model construction of human serum protein binding (Votano et al. 2006). A robust ANN model based of topological descriptors in combination with logP was established with $r^2 = 0.90$, MAE = 7.6 and $r^2 = 0.70$, MAE = 14.1 respectively. MAE stands for Mean Absolute Error.

Votano's data set was used by Ghafourian et al. (Ghafourian and Amin 2013) to construct linear regression and nonlinear models using classification and regression trees (CART), boosted trees and random forest. Interpretable linear regression and simple regression trees models were able to identify the important contribution of hydrophobicity, van der Waals surface area and aromaticity for high PPB. On the other hand, the more complicated ensemble method of boosted regression trees produced the most accurate PPB predictions.

Combination of chemometrics with molecular modeling confirmed the preponderant contribution of hydrophobic regions of drug molecules and the specific roles of polar groups, which anchor drugs to HSA 1 and 2 binding sites (Estrada et al. 2006). Identification of the binding site before performing QSAR analysis can evidently lead to better models. For 889 chemically diverse compounds with binding affinity for domain III-A, a group contribution model was developed based on 74 chemical fragments. ($R^2 = 0.94$, $Q^2 = 0.90$) (Hajduk et al. 2003).

The authors further suggested a combination of QSAR models for full-length albumin and for domain-IIIa to allow for discrimination between compounds that bind to the latter site and those that bind elsewhere on the protein. An important issue is that the fragments used in the model are mapped by most of the topological descriptors included in Votano's model, indicating that they can be considered quite universal. Thus, they provide a convenient look-up table for quantitatively estimation of the effect of a particular group to albumin binding.

A free web prediction platform was constructed by Zsila et al. who combined support vector machine (SVM) classification model with molecular docking calculations. The classification model was based on 45 descriptors, with logP being the most important. The platform (<http://albumin.althotas.com>) enables the users (i) to predict if albumin binds the query ligand, (ii) to determine the probable ligand binding site (site 1 or site 2) according to the classification model (iii) to select using the Tanimoto similarity the albumin X-ray structure which is complexed with the most similar ligand and (iv) to calculate complex geometry using molecular docking calculations (Zsila 2013).

The continuous update of the HSA models in order to maintain their performance over time is essential for the drug discovery and development settings, extending their applicability domain and robustness. In this sense, Rodgers et al. proposed a procedure for monthly updating human plasma protein binding models over a period of 21 months (Rodgers et al. 2007), which was extended to three years, using partial least squares (PLS), random forest (RF) and Bayesian neural networks (BNN). The authors started with a large data set, the size of which was doubled by the end of the study (Rodgers et al. 2011). Consensus predictions of HSA binding constants using the final models, generated by all three techniques showed, RMSE = 0.55. These results justified the need for the automatic regular updating of QSAR models (autoQSAR) in the case of ADME properties.

An analogous approach for modeling HSA binding, as well as other ADME properties, over time is implemented in a software architecture, the so called "Discovery Bus" which allows exhaustive exploration of descriptor and model space, automates model validation and their continuous updating providing an automated QSPR through competitive workflow (Cartmell et al. 2005).

Recently, ensemble machine learning-based QSPR models have been established for a four-category classification and PPB affinity prediction, using a dataset of 930 compounds. The structural diversity of the compounds was tested by the Tanimoto similarity index. In the test set, the classification QSPR models proved superior with an accuracy > 93%, while the regression QSPR models yielded $r^2 > 0.920$ between the measured and predicted PPB affinities, with the root mean squared error < 9.77. Lipophilicity, expressed as XLogP, was the most important descriptor (Basant et al. 2016).

For further PPB models and for the state of the art in predicting binding to α 1-acid glycoprotein, the second important plasma protein, the reader is referred to a recent comprehensive review by Lambrinidis et al. (2015).

5.2.5 Prediction Models for Metabolism

Metabolism, the M in 'ADME', is one of the main factors influencing the fate and toxicity of a chemical. Metabolism or (biotransformation) includes a large set of chemical reactions, which generally convert drugs or other xenobiotics into more polar and more easily excreted, i.e., less toxic forms. However, in some cases, metabolism may lead to toxic metabolites or/and intermediates. Thus, metabolites with physicochemical and pharmacological properties that differ substantially from those of the parent drug have important implications for both drug safety and efficacy (Testa et al. 2004; Testa 2009).

The utility of conventional QSARs predicting the metabolic fate of chemicals is rather limited. Most of the models are established to predict the phase I metabolism, mainly addressing cytochrome P450 (CYP450) isoforms, a superfamily of enzymes including more than 70 families of proteins, which play a predominant role in the biotransformation of drugs and xenobiotics. Based on a 'guesstimate' of the number of drug metabolites that are known to be produced by cytochromes P450 isoforms and other oxidoreductases (EC 1), as well as hydrolases (EC 3), and transferases (EC 2), it is supposed that oxidoreductases are the main enzymes responsible for the formation of toxic or active metabolites, whereas transferases play the major role in producing inactive and nontoxic metabolites (Testa 2009).

Terfloth et al. (2007) investigated the application of several model-building techniques, such as k-NN, decision trees, Multilayer Perceptron as Neural Networks (MLPNN), Radial Basis Function Neural Networks (RBF-NN), Logistic Regression (LR) and Support Vector Machine (SVM), to predict the isoform specificity for CYP450 3A4, 2D6 and 2C9 substrates (Terfloth et al. 2007). The applied descriptors included simple molecular properties and functional group accounts, topological descriptors, descriptors related to the shape of molecules or the distribution of interatomic distances considering the 3D structures of the molecules. A 9-descriptor model, established by combining automatic variable selection with the SVM technique, gave the best results. The achieved predictivity for an external data set of 233 compounds was equal 83%. Promising results were also obtained for the decision tree based model with three descriptors only, and 80% predictivity for the external data set was achieved. Burton et al. (2006) constructed classification models for human CYP1A2 and CYP2D6 inhibition using binary decision tree. The decision tree for CYP2D6 had sensitivity 88%, specificity 92% and positive predictivity 90%. The external validation had a accuracy 89%, sensitivity 91%, specificity 92% and precision 90%. For CYP1A2, accuracy was 89%, sensitivity 95%, specificity 83% and precision 85% for the training set while the test set had 81% accuracy, 76% sensitivity, 86% specificity and 85% precision. The authors identified a range of useful descriptors. Van der Waals surface area (VSA) was particularly efficient and allowed to develop models reaching 95% correct classification. 3D descriptors also provided promising results. Sheridan et al. (2007) applied Random Forest (RF) technique for predicting CYP450 (3A4, 2D6, 2C9) sites of the metabolism, using descriptors that describe the environment around each non-hydrogen atom in each molecule. The authors identified several descriptors positively and

negatively related to the oxidation sites of molecules. Compared to the results using MetaSite software (Molecular Discovery) of Cruciani et al. (2005), Sheridan's model performed better in the case of CYP3A4. For CYP2D6 and CYP2C9 the predictions of Sheridan's model were only slightly better.

In the case of metabolism, computer-based expert systems have a much broader applicability. Among them MetaSite is widely used (Cruciani et al. 2005). It makes predictions based on the lability of hydrogens and orientation effects derived from the 3D structure of a CYP active site, independently of the availability of pre-existing data. MetaSite can handle 3A4, 2D6, 2C9, 1A2, 2C9, and 2C19 and can be extended to any CYP for which a homology model can be generated. It is advantageous for enzymes such as CYP1A2 and CYP2C19, where there are not currently enough data in the literature to generate a QSAR model. Moreover, the MetaSite methodology is easy to use, fast and fully automated. Other expert systems are MetabolExpert, developed by CompuDrug (Darvas 1988), METEOR (Testa et al. 2005a) COMPACT (Computer-Optimised Molecular Parametric Analysis of Chemical Toxicity) (Lewis et al. 1996; Lewis 2001) and META, implemented in MCASE ADME Module (MultiCASE) (Klopman et al. 1999, 1997; Talafous et al. 1994).

More information about for predicting drug metabolism can be found in a recent review by Kirchmair et al. (2015).

5.2.6 Integrated ADME Prediction Models

In previous sections, separate models for different processes along the drug discovery and development pipeline are discussed. The medicinal chemist team should try to take advantage by applying them in their project compounds, selected by early stage techniques, e.g., virtual screening, structure or ligand based design for the target of interest, drug-like filtering. The multi-objective character of drug development however has challenged the creation of software tools and web platforms mainly for the purpose of integrated ADME and ADME-related predictions. Many of them are commercial. They differ greatly in terms of their capabilities and applications. Prediction software for physicochemical properties like lipophilicity and ionization, related to ADME, has already been discussed in Sect. 4. Solubility is another endpoint of interest for oral absorption as well as for formulation issues. Such predictions serve as inputs to models of key ADME properties, mainly for gastrointestinal absorption, BBB permeability, oral bioavailability (including affinity to uptake or efflux transporter) and plasma protein binding. Predictions of possible metabolite, as well as toxicity endpoints like mutagenicity, carcinogenicity or teratogenicity are also implemented in certain software. Some popular software are Know-it-All (Bio-Rad Laboratories <http://www.bio-rad.com/>), ADME Boxes (Pharma Algorithms—now included in ACD/ADME Suite), and ADMET Predictor (Simulations Plus Inc. <http://www.simulations-plus.com/>). VolSurf/VolSurf + (Molecular Discovery and Tripos) also predicts various ADME properties including passive intestinal absorption,

blood-brain barrier permeation, solubility, protein binding, volume of distribution, and metabolic stability on the basis of different models based on VolSurf descriptors.

Moreover, there is a trend towards developing more sophisticated, mathematical PBPK models, see also Sect. 5.2.2. In these software tools, *in vitro* and/or *in vivo* ADME data are integrated with the results of QSAR/QSPR models (e.g., for percentage plasma protein binding or blood/brain barrier penetration) for organism-based ADME modeling. GastroPlus and Cloe, which mimic the processes inside living organisms, are more commonly used. Simcyp (<http://www.simcyp.com/>) is a proprietary PBPK simulator that provides a platform for modeling the ADME properties of drugs and their metabolites, as well as drug-drug interactions, in virtual patient populations (Jamei et al. 2009).

It should be noted as a warning for using software for ADME prediction that the results should be considered as rough estimates, useful for screening purposes or as starting points for further modeling or experimental evidence.

6 Conclusions

Drug discovery and development is a complicated multi-objective and expensive enterprise, with drug candidates being a compromise of competing pharmacodynamics and pharmacokinetic processes. *In silico* predictions along the different stages of the pipeline provide valuable support in the selection of drug candidates with balanced properties, so as to control each stage early enough and reduce failures at clinical phases. High technology provides new endpoints that may serve to establish efficient QSAR and QSPR models, which themselves profit of the evolution in computational and statistical techniques. Local and global models have their own value, dependent on the underlying goal and the timeline. Initial screening, off-target affinities or ADME properties benefit more by global models, while local models are suitable for selected project ligands with potential affinity for a target receptor. Interpretability of models is an important issue. The medicinal chemist is more familiar with models containing well understandable physico-chemical or molecular descriptors, which provide an insight in the mechanism of action. However the most accurate model is not always the most interpretable. In such cases the intended use of the model is the determinant factor. Nevertheless, toxicity models for regulatory purpose must have a certain degree of interpretability as required by OECD.

The correct use of the models implies that the user is aware of their merits and pitfalls. Their evaluation should consider the accuracy and range of the endpoints, while external validation with blind test sets is a strict prerequisite in particular for global models. In such cases, determination of their applicability is useful in order to evaluate when predictions are reliable.

In conclusion, the results of the *in silico* models at the different stages of drug discovery should be taken into consideration for prioritizing the drug candidates,

before proceeding to the next step. The ultimate goal is to produce safe and efficient drug candidates, a goal, which can be achieved by finding the golden ratio between affinity to the target receptor, in regard also to off-targets and the appropriate pharmacokinetic properties in compliance with the concept of druglikeness. The tools are available, they need to be properly used.

References

- Abad-Zapatero, C. (2007). Ligand efficiency indices for effective drug discovery. *Expert Opinion Drug Discovery*, 2, 469–488. doi:10.1517/17460441.2.4.469.
- Abraham, M. H., Ibrahim, A., Zhao, Y., & Acree, W. E. (2006). A data base for partition of volatile organic compounds and drugs from blood/plasma/serum to brain, and an LFER analysis of the data. *Journal of Pharmaceutical Sciences*, 95, 2091–2100.
- Abraham, M. H., Ibrahim, A., Zissimos, A. M., Zhao, Y. H., Comer, J., & Reynolds, D. P. (2002). Application of hydrogen bonding calculations in property based drug design. *Drug Discovery Today*, 7, 1056–1063.
- Akhondi, S. A., Kors, J. A., & Muresan, S. (2012). Consistency of systematic chemical identifiers within and between small-molecule databases. *Journal of Cheminformatics*, 4, 1.
- Anderson, A. C. (2003). The process of structure-based drug design. *Chemistry & Biology*, 10, 787–797.
- Andrews, C. W., Bennett, L., & Lawrence, X. Y. (2000). Predicting human oral bioavailability of a compound: Development of a novel quantitative structure-bioavailability relationship. *Pharmaceutical Research*, 17, 639–644.
- Artursson, P., Palm, K., & Luthman, K. (2001). Caco-2 monolayers in experimental and theoretical predictions of drug transport. *Advanced Drug Delivery Reviews*, 46, 27–43.
- Ashour, M.-B. A., Gee, S. J., & Hammock, B. D. (1987). Use of a 96-well microplate reader for measuring routine enzyme activities. *Analytical Biochemistry*, 166, 353–360.
- Avdeef, A. (2012). *Absorption and drug development: solubility, permeability, and charge state*. Wiley.
- Balaban, A. T. (Ed.). (1997). *From chemical topology to three-dimensional geometry*. New York (NY): Plenum Press.
- Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., et al. (2007). NCBI GEO: Mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Research*, 35, D760–D765.
- Basant, N., Gupta, S., & Singh, K. P. (2016). Predicting binding affinities of diverse pharmaceutical chemicals to human serum plasma proteins using QSPR modelling approaches. *SAR and QSAR in Environmental Research*, 27, 67–85.
- Bergström, C. A., Charman, W. N., & Porter, C. J. (2016). Computational prediction of formulation strategies for beyond-rule-of-5 compounds. *Advanced Drug Delivery Reviews*, 101, 6–21.
- Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., & Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4, 90–98.
- Birchall, K., Gillet, V. J., Harper, G., & Pickett, S. D. (2008a). Evolving interpretable Structure-activity relationships. 1. Reduced graph queries. *Journal of Chemical Information and Modeling*, 48, 1543–1557.
- Birchall, K., Gillet, V. J., Harper, G., & Pickett, S. D. (2008b). Evolving interpretable structure-activity relationship models. 2. Using multiobjective optimization to derive multiple models. *Journal of Chemical Information and Modeling*, 48, 1558–1570.
- Bois, F. Y., & Brochot, C. (2016). Modeling pharmacokinetics. In E. Benfenati (Ed.), *Silico Methods for Predicting Drug Toxicity* (pp. 37–62). New York, NY: Springer New York.

- Brandt, T., Holzmann, N., Muley, L., Khayat, M., Wegscheid-Gerlach, C., Baum, B., et al. (2011). Congeneric but still distinct: How closely related trypsin ligands exhibit different thermodynamic and structural properties. *Journal of Molecular Biology*, 405, 1170–1187. doi:10.1016/j.jmb.2010.11.038.
- Brown, A. C., & Fraser, T. R. (1868). On the connection between chemical constitution and physiological action; with special reference to the physiological action of the salts of the ammonium bases derived from strychnia, brucia, thebaia, codeia, morphia, and nicotia. *Journal of Anatomy and Physiology*, 2, 224.
- Bruce, C. L., Melville, J. L., Pickett, S. D., & Hirst, J. D. (2007). Contemporary QSAR classifiers compared. *Journal of Chemical Information and Modeling*, 47, 219–227.
- Burton, J., Ijjaali, I., Barberan, O., Petitet, F., Vercauteren, D. P., & Michel, A. (2006). Recursive partitioning for the prediction of cytochromes P450 2D6 and 1A2 inhibition: importance of the quality of the dataset. *Journal of Medicinal Chemistry*, 49, 6231–6240.
- Byvatov, E., Fechner, U., Sadowski, J., & Schneider, G. (2003). Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of Chemical Information and Computer Sciences*, 43, 1882–1889.
- Cabrera-Perez, M. A., Bermejo, M., Alvarez, I. G., Alvarez, M. G., Garrigues, T. M., et al. (2012). QSPR in oral bioavailability: Specificity or integrality? *Mini Reviews in Medicinal Chemistry* 12, 534–550.
- Caporuscio, F., & Tafi, A. (2011). Pharmacophore modelling: A forty year old approach and its modern synergies. *Current Medicinal Chemistry*, 18, 2543–2553.
- Cartmell, J., Enoch, S., Krstajic, D., & Leahy, D. E. (2005). Automated QSPR through competitive workflow. *Journal of Computer-Aided Molecular Design*, 19, 821–833.
- Castillo-Garit, J. A., Marrero-Ponce, Y., Torrens, F., & García-Domenech, R. (2008). Estimation of ADME properties in drug discovery: Predicting Caco-2 cell permeability using atom-based stochastic and non-stochastic linear indices. *Journal of Pharmaceutical Sciences*, 97, 1946–1976.
- Cereto-Massagué, A., Guasch, L., Valls, C., Mulero, M., Pujadas, G., & Garcia-Vallvé, S. (2012). DecoyFinder: an easy-to-use python GUI application for building target-specific decoy sets. *Bioinformatics*, 28, 1661–1662.
- Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I.I., Cronin, M., et al. (2014). QSAR modeling: where have you been? Where are you going to? *Journal of Medicinal Chemistry* 57, 4977–5010.
- Chrysanthakopoulos, M., Koletsou, A., Nicolaou, I., Demopoulos, V. J., & Tsantili-Kakoulidou, A. (2009). Lipophilicity studies on pyrrolyl-acetic acid derivatives. Experimental versus predicted logP values in relationship with aldose reductase inhibitory activity. *QSAR & Combinatorial Science*, 28, 551–560.
- Clark, D. E. (2003). In silico prediction of blood–brain barrier permeation. *Drug Discovery Today*, 8, 927–933.
- Clark, D. E. (1999). Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *Journal of Pharmaceutical Sciences*, 88, 807–814.
- Congreve, M., Carr, R., Murray, C., & Jhoti, H. (2003). A “rule of three” for fragment-based lead discovery? *Drug Discovery Today* 8, 876–877.
- Consonni, V., Todeschini, R., & Pavan, M. (2002). Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *Journal of Chemical Information and Computer Sciences*, 42, 682–692.
- Cox, R., Green, D. V., Luscombe, C. N., Malcolm, N., & Pickett, S. D. (2013). QSAR workbench: Automating QSAR modeling to drive compound design. *Journal of Computer-Aided Molecular Design*, 27, 321–336.
- Cramer, R. D., Patterson, D. E., & Bunce, J. D. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society*, 110, 5959–5967.

- Cramer, R. D. (2012). The inevitable QSAR renaissance. *Journal of Computer-Aided Molecular Design*, 26, 35–38.
- Crivori, P., Cruciani, G., Carrupt, P.-A., & Testa, B. (2000). Predicting blood-brain barrier permeation from three-dimensional molecular structure. *Journal of Medicinal Chemistry*, 43, 2204–2216.
- Cruciani, G., Carosati, E., De Boeck, B., Ethirajulu, K., Mackie, C., Howe, T., et al. (2005). MetaSite: Understanding metabolism in human cytochromes from the perspective of the chemist. *Journal of Medicinal Chemistry*, 48, 6970–6979.
- Cruciani, G., Crivori, P., Carrupt, P.-A., & Testa, B. (2000). Molecular fields in quantitative structure—Permeation relationships: The VolSurf approach. *Journal of Molecular Structure: THEOCHEM*, 503, 17–30.
- Csizmadia, F., Tsantili-Kakoulidou, A., Panderi, I., & Darvas, F. (1997). Prediction of distribution coefficient from structure. 1. Estimation method. *Journal of Pharmaceutical Sciences*, 86, 865–871.
- Darvas, F. (1988). Predicting metabolic pathways by logic programming. *Journal of Molecular Graphics*, 6, 80–86.
- Dearden, J. C. (2007). In silico prediction of ADMET properties: How far have we come? *Expert Opinion Drug Metabolism Toxicology*, 3, 635–639.
- De Benedetti, P. G., & Fanelli, F. (2010). Computational quantum chemistry and adaptive ligand modeling in mechanistic QSAR. *Drug Discovery Today*, 15, 859–866.
- De Melo, E. B., Ferreira, M. M. C., et al. (2009). Nonequivalent effects of diverse LogP algorithms in three QSAR studies. *QSAR Comb Sci* 28, 1156–1165.
- Di, L., Kerns, E. H., Bezar, I. F., Petusky, S. L., & Huang, Y. (2009). Comparison of blood–brain barrier permeability assays: in situ brain perfusion, MDR1-MDCKII and PAMPA-BBB. *Journal of Pharmaceutical Sciences*, 98, 1980–1991.
- Dunn, W. J. (1988). QSAR approaches to predicting toxicity. *Toxicology Letters*, 43, 277–283.
- Ecker, G. F., & Noe, C. R. (2004). In silico prediction models for blood–brain barrier permeation. *Current Medicinal Chemistry*, 11, 1617.
- Ekins, S., Bravi, G., Binkley, S., Gillespie, J. S., Ring, B. J., Wikel, J. H., et al. (1999). Three- and four-dimensional quantitative structure activity relationship analyses of cytochrome P-450 3A4 inhibitors. *Journal of Pharmacology and Experimental Therapeutics*, 290, 429–438.
- Enyedy, I. J., & Egan, W. J. (2008). Can we use docking and scoring for hit-to-lead optimization? *Journal of Computer-Aided Molecular Design*, 22, 161–168. doi:10.1007/s10822-007-9165-4.
- Enyedy, I. J., Ling, Y., Nacro, K., Tomita, Y., Wu, X., Cao, Y., et al. (2001). Discovery of small-molecule inhibitors of Bcl-2 through structure-based computer screening. *Journal of Medicinal Chemistry*, 44, 4313–4324.
- Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wold, S. (2001). *Multi- and megavariate data analysis: Principles and applications. Umetrics*.
- Estrada, E., Uriarte, E., Molina, E., Simón-Manso, Y., & Milne, G. W. (2006). An integrated in silico analysis of drug-binding to human serum albumin. *Journal of Chemical Information and Modeling*, 46, 2709–2724.
- Evans, W. E., & Guy, R. K. (2004). Gene expression as a drug discovery tool. *Nature Genetics*, 36, 214–215.
- Faulon, J.-L., Brown, W. M., & Martin, S. (2005). Reverse engineering chemical structures from molecular descriptors: how many solutions? *Journal of Computer-Aided Molecular Design*, 19, 637–650.
- Filikov, A. V., Mohan, V., Vickers, T. A., Griffey, R. H., Cook, P. D., Abagyan, R. A., et al. (2000). Identification of ligands for RNA targets via structure-based virtual screening: HIV-1 TAR. *Journal of Computer-Aided Molecular Design*, 14, 593–610.
- Fujita, T., & Winkler, D. A. (2016). Understanding the roles of the “two QSARs”. *Journal of Chemical Information and Modeling*, 56, 269–274.
- Ganellin, C. R. (2004). Robin Ganellin gives his views on medicinal chemistry and drug discovery. *Drug Discovery Today*, 9, 158–160.
- Gasteiger, J., et al. (2003). *Handbook of chemoinformatics*. Wiley Online Library.

- Gaviraghi, G., Barnaby, R. J., & Pellegatti, M. (2001). Pharmacokinetic challenges in lead optimization. *Testa B Van Waterbeemd H folk. G* 3–14.
- Gedeck, P., Rohde, B., & Bartels, C. (2006). QSAR-how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. *Journal of Chemical Information and Modeling*, 46, 1924–1936.
- Ghafourian, T., & Amin, Z. (2013). QSAR models for the prediction of plasma protein binding. *BiolImpacts BI*, 3, 21.
- Giaginis, C., Theocharis, S., & Tsantili-Kakoulidou, A. (2008). Quantitative Structure-activity relationships for PPAR- γ binding and gene transactivation of tyrosine-based agonists using multivariate statistics. *Chemical Biology & Drug Design*, 72, 257–264.
- Giaginis, C., Theocharis, S., & Tsantili-Kakoulidou, A. (2007). A consideration of PPAR- γ ligands with respect to lipophilicity: Current trends and perspectives. *Expert Opinion on Investigational Drugs*, 16, 413–417.
- Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., & Chong, J. (2016). BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, 44, D1045–D1053.
- Godden, J. W., & Bajorath, J. (2000). Shannon entropy—A novel concept in molecular descriptor and diversity analysis. *Journal of Molecular Graphics and Modelling*, 18, 73–76.
- Goldmann, D., Montanari, F., Richter, L., Zdrzil, B., & Ecker, G. F. (2014). Exploiting open data: A new era in pharmacoinformatics. *Future Medicinal Chemistry*, 6, 503–514.
- Goodford, P. J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry*, 28, 849–857.
- Gozalbes, R., Doucet, J. P., & Derouin, F. (2002). Application of Topological descriptors in QSAR and drug design: History and new trends. *Current Drug Targets-Infectious Disorders*, 2, 93–102. doi:10.2174/1568005024605909.
- Gramatica, P. (2006). WHIM descriptors of shape. *QSAR & Comb. Sci.*, 25, 301–415.
- Guha, R. (2011). The ups and downs of structure-activity landscapes. *Chemoinformatics and Computational Chemical Biology*, 101–117.
- Guha, R., & Jurs, P. C. (2005). Determining the validity of a QSAR model—a classification approach. *Journal of Chemical Information and Modeling*, 45, 65–73.
- Hajduk, P. J., Mendoza, R., Petros, A. M., Huth, J. R., Bures, M., Fesik, S. W., et al. (2003). Ligand binding to domain-3 of human serum albumin: A chemometric analysis. *Journal of Computer-Aided Molecular Design*, 17, 93–102.
- Hann, M. M. (2011). Molecular obesity, potency and other addictions in drug discovery. *MedChemComm*, 2, 349–355.
- Hansch, C. (1969). Quantitative approach to biochemical structure-activity relationships. *Accounts of Chemical Research*, 2, 232–239.
- Hansch, C., & Clayton, J. M. (1973). Lipophilic character and biological activity of drugs II: The parabolic case. *Journal of Pharmaceutical Sciences*, 62, 1–21.
- Hansch, C., & Fujita, T. (1964). ρ - σ - π Analysis. A method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, 86, 1616–1626.
- Hansch, C., Hoekman, D., Leo, A., Zhang, L., & Li, P. (1995a). The expanding role of quantitative structure-activity relationships (QSAR) in toxicology. *Toxicology Letters*, 79, 45–53.
- Hansch, C., & Leo, A. (1979). Substituent constants for correlation analysis in chemistry and biology. Wiley.
- Hansch, C., Leo, A., Hoekman, D. H., et al. (1995b). Exploring QSAR: Fundamentals and applications in chemistry and biology. Washington, DC: American Chemical Society.
- Hansch, C., Muir, R. M., Fujita, T., Maloney, P. P., Geiger, F., & Streich, M. (1963). The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients. *Journal of the American Chemical Society*, 85, 2817–2824.
- Hanumegowda, U. M., Wenke, G., Regueiro-Ren, A., Yordanova, R., Corradi, J. P., & Adams, S. P. (2010). Phospholipidosis as a function of basicity, lipophilicity, and volume of distribution of compounds. *Chemical Research in Toxicology*, 23, 749–755.

- Helgee, E. A., Carlsson, L., Boyer, S., & Norinder, U. (2010). Evaluation of quantitative structure-activity relationship modeling strategies: Local and global models. *Journal of Chemical Information and Modeling*, *50*, 677–689.
- Helguera A. M., Combes R. D., González M. P., & Cordeiro M. N. (2008). Applications of 2D descriptors in drug design: A DRAGON tale. *Current Topics in Medicinal Chemistry*, *8*, 1628–1655.
- Hertzberg, R. P., & Pope, A. J. (2000). High-throughput screening: New technology for the 21st century. *Current Opinion in Chemical Biology*, *4*, 445–451.
- Hieronimus, H., Lamb, J., Ross, K. N., Peng, X.P., Clement, C., Rodina, A., et al. (2006). Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell* *10*, 321–330.
- Hillisch, A., Heinrich, N., & Wild, H. (2015). Computational chemistry in the pharmaceutical industry: From childhood to adolescence. *ChemMedChem*, *10*, 1958–1962.
- Hodos, R. A., Kidd, B. A., Shameer, K., Readhead, B. P., & Dudley, J. T. (2016). In silico methods for drug repurposing and pharmacology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *8*, 186–210.
- Hollósy, F., Valkó, K., Hersey, A., Nunhuck, S., Kéri, G., & Bevan, C. (2006). Estimation of volume of distribution in humans from high throughput HPLC-based measurements of human serum albumin binding and immobilized artificial membrane partitioning. *Journal of Medicinal Chemistry*, *49*, 6958–6971.
- Hopfinger, A. J., Wang, S., Tokarski, J. S., Jin, B., Albuquerque, M., Madhav, P. J., et al. (1997). Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *Journal of the American Chemical Society*, *119*, 10509–10524.
- Hopkins, A. L., Keserü, G. M., Leeson, P. D., Rees, D. C., & Reynolds, C. H. (2014). The role of ligand efficiency metrics in drug discovery. *Nature Reviews Drug Discovery*, *13*, 105–121.
- Houston, J. G., & Banks, M. (1997). The chemical-biological interface: Developments in automated and miniaturised screening technology. *Current Opinion in Biotechnology*, *8*, 734–740.
- Hou, T., Wang, J., Zhang, W., & Xu, X. (2007). ADME evaluation in drug discovery. 6. Can oral bioavailability in humans be effectively predicted by simple molecular property-based rules? *Journal of Chemical Information and Modeling*, *47*, 460–463.
- Huang, S.-M., Abernethy, D. R., Wang, Y., Zhao, P., & Zineh, I. (2013). The utility of modeling and simulation in drug development and regulatory review. *Journal of Pharmaceutical Sciences*, *102*, 2912–2923.
- Hughes, J. D., Blagg, J., Price, D. A., Bailey, S., DeCrescenzo, G. A., Devraj, R. V., et al. (2008). Physicochemical drug properties associated with in vivo toxicological outcomes. *Bioorganic & Medicinal Chemistry Letters* *18*, 4872–4875.
- Irvine, J. D., Takahashi, L., Lockhart, K., Cheong, J., Tolan, J. W., Selick, H. E., et al. (1999). MDCK (Madin–Darby canine kidney) cells: A tool for membrane permeability screening. *Journal of Pharmaceutical Sciences*, *88*, 28–33.
- Irwin, J. J. (2008). Using ZINC to acquire a virtual screening library. *Current Protocols in Bioinformatics*, 14–6.
- Ito, K., Iwatsubo, T., Kanamitsu, S., Nakajima, Y., & Sugiyama, Y. (1998). Quantitative prediction of in vivo drug clearance and drug interactions from in vitro data on metabolism, together with binding and transport. *Annual Review of Pharmacology and Toxicology*, *38*, 461–499. doi:10.1146/annurev.pharmtox.38.1.461.
- Jamei, M. (2016). Recent advances in development and application of physiologically-based pharmacokinetic (PBPK) models: A transition from academic curiosity to regulatory acceptance. *Current Pharmacology Reports*, *2*, 161–169.
- Jamei, M., Marciniak, S., Feng, K., Barnett, A., Tucker, G., & Rostami-Hodjegan, A. (2009). The Simcyp® population-based ADME simulator. *Expert Opinion on Drug Metabolism & Toxicology*, *5*, 211–223.
- Jaworska, J., Nikolova-Jeliazkova, N., & Aldenberg, T. (2005). QSAR applicability domain estimation by projection of the training set descriptor space: A review. *Atla-Nottingham* *33*, 445.

- Jaworska, J. S., Comber, M., Auer, C., & Van Leeuwen, C. J. (2003). Summary of a workshop on regulatory acceptance of (Q) SARs for human health and environmental endpoints. *Environmental Health Perspectives*, *111*, 1358.
- Jorgensen, W. L. (2009). Efficient drug lead discovery and optimization. *Accounts of Chemical Research*, *42*, 724–733.
- Jorgensen, W. L. (2004). The many roles of computation in drug discovery. *Science*, *303*, 1813–1818.
- Kaldor, S. W., Kalish, V. J., Davies, J. F., Shetty, B.V., Fritz, J.E., Appelt, K., et al. (1997). Viracept (nelfinavir mesylate, AG1343): A potent, orally bioavailable inhibitor of HIV-1 protease. *Journal of Medicinal Chemistry*, *40*, 3979–3985.
- Kaliszan, R., & Markuszewski, M. (1996). Brain/blood distribution described by a combination of partition coefficient and molecular mass. *International Journal of Pharmaceutics*, *145*, 9–16.
- Kansy, M., Senner, F., & Gubernator, K. (1998). Physicochemical high throughput screening: Parallel artificial membrane permeation assay in the description of passive absorption processes. *Journal of Medicinal Chemistry*, *41*, 1007–1010.
- Kariv, I., Cao, H., & Oldenburg, K. R. (2001). Development of a high throughput equilibrium dialysis method. *Journal of Pharmaceutical Sciences*, *90*, 580–587.
- Katritzky, A. R., Lobanov, V. S., & Karelson, M. (1994). *CODESSA: reference manual*. FL: Univ. Fla. Gainesv.
- Keserü, G. M. (2001). A virtual high throughput screen for high affinity cytochrome P450cam substrates. Implications for in silico prediction of drug metabolism. *Journal of Computer-Aided Molecular Design*, *15*, 649–657.
- Kier, L. B., & Hall, L. H. (1999). *Molecular structure description: The Electrotopological State*. San Diego, CA: Academic Press.
- Kim, M. T., Sedykh, A., Chakravarti, S. K., Saiakhov, R. D., & Zhu, H. (2014). Critical evaluation of human oral bioavailability for pharmaceutical drugs by using various cheminformatics approaches. *Pharmaceutical Research*, *31*, 1002–1014.
- Kirchmair, J., Göller, A. H., Lang, D., Kunze, J., Testa, B., Wilson, I. D., et al. (2015). Predicting drug metabolism: experiment and/or computation? *Nature Reviews Drug Discovery*, *14*, 387–404.
- Klebe, G. (1998). Comparative molecular similarity indices analysis: CoMSIA. *Perspectives in Drug Discovery and Design*, *12*, 87–104.
- Klebe, G., Abraham, U., & Mietzner, T. (1994). Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *Journal of Medicinal Chemistry*, *37*, 4130–4146.
- Klopman, G., & Kalos, A. N. (1985). Causality in structure-activity studies. *Journal of Computational Chemistry*, *6*, 492–506.
- Klopman, G., Stefan, L. R., & Saiakhov, R. D. (2002). ADME evaluation: 2. A computer model for the prediction of intestinal absorption in humans. *European Journal of Pharmaceutical Sciences*, *17*, 253–263.
- Klopman, G., Tu, M., & Fan, B. T. (1999). META 4. Prediction of the metabolism of polycyclic aromatic hydrocarbons. *Theoretical Chemistry Accounts*, *102*, 33–38.
- Klopman, G., Tu, M., & Talafous, J. (1997). META. 3. A genetic algorithm for metabolic transform priorities optimization. *Journal of Chemical Information and Computer Sciences*, *37*, 329–334.
- Kola, I., & Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery*, *3*, 711–715. doi:10.1038/nrd1470.
- Kononov, D. A., Coomans, D., Deconinck, E., & Vander Heyden, Y. (2007). Benchmarking of QSAR models for blood-brain barrier permeation. *Journal of Chemical Information and Modeling*, *47*, 1648–1656.
- Koukoulitsa, C., Tsantili-Kakoulidou, A., Mavromoustakos, Th., & Chinou, I. (2009). PLS analysis for antibacterial Activity of natural coumarins using Volsurf descriptors. *QSAR and Comb. Sci.*, *28*, 785–789.

- Kubinyi, H. (1979). Lipophilicity and drug activity, in: *Progress in Drug Research/Fortschritte Der Arzneimittelforschung/Progress Des Recherches Pharmaceutiques*, pp. 97–198. Springer.
- Kubinyi, H., & Kehrhahn, O. H. (1978). Quantitative structure-activity relationships. VI. Non-linear dependence of biological activity on hydrophobic character: Calculation procedures for bilinear model. *Arzneimittel-Forschung*, 28, 598–601.
- Kubinyi, H., Mannhold, R., Krogsgaard, L. R., & Timmerman, H. E., (1993). In R. Mannhold, Al (Eds.), *Methods and principles in medicinal chemistry*.
- Kumar, R., Sharma, A., & Varadwaj, P. K. (2011). A prediction model for oral bioavailability of drugs using physicochemical properties by support vector machine. *Journal of Natural Science, Biology, and Medicine*, 2, 168.
- Lambrinidis, G., Vallianatou, T., & Tsantili-Kakoulidou, A. (2015). In vitro, in silico and integrated strategies for the estimation of plasma protein binding. A review. *Advanced Drug Delivery Reviews*, 86, 27–45.
- Larregieu, C. A., & Benet, L. Z. (2014). Distinguishing between the permeability relationships with absorption and metabolism to improve BCS and BDDCS predictions in early drug discovery. *Molecular Pharmaceutics*, 11, 1335–1344.
- Larregieu, C. A., & Benet, L. Z. (2013). Drug discovery and regulatory considerations for improving in silico and in vitro predictions that use Caco-2 as a surrogate for human intestinal permeability measurements. *American Association of Pharmaceutical Scientists Journal*, 15, 483–497.
- Leeson, P. D., & Springthorpe, B. (2007). The influence of drug-like concepts on decision-making in medicinal chemistry. *Nature Reviews Drug Discovery*, 6, 881–890.
- Leo, A., Hansch, C., & Elkins, D. (1971). Partition coefficients and their uses. *Chemical Reviews*, 71, 525–616. doi:10.1021/cr60274a001.
- Levin, V. A. (1980). Relationship of octanol/water partition coefficient and molecular weight to rat brain capillary permeability. *Journal of Medicinal Chemistry*, 23, 682–684.
- Lewis, D. F., Ioannides, C., & Parke, D. V. (1996). COMPACT and molecular structure in toxicity assessment: A prospective evaluation of 30 chemicals currently being tested for rodent carcinogenicity by the NCI/NTP. *Environmental Health Perspectives*, 104, 1011.
- Lewis, D. F. V. (2001). COMPACT: A structural approach to the modelling of cytochromes P450 and their interactions with xenobiotics. *Journal of Chemical Technology and Biotechnology*, 76, 237–244.
- Lima, A. N., Philot, E. A., Trossini, G. H. G., Scott, L. P. B., Maltarollo, V. G., & Honorio, K. M. (2016). Use of machine learning approaches for novel drug discovery. *Expert Opinion on Drug Discovery*, 11, 225–239.
- Lind, K. E., Du, Z., Fujinaga, K., Peterlin, B. M., & James, T. L. (2002). Structure-based computational database screening, in vitro assay, and NMR assessment of compounds that target TAR RNA. *Chemistry & Biology*, 9, 185–193.
- Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (1997). In Vitro Models for Selection of Development Candidates: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23, 3–25. doi:10.1016/S0169-409X(96)00423-1.
- Löfås, S., & Johansson, B. (1990). A novel hydrogel matrix on gold surfaces in surface plasmon resonance sensors for fast and efficient covalent immobilization of ligands. *Journal of the Chemical Society, Chemical Communications*, 1526–1528.
- Luco, J. M. (1999). Prediction of the brain-blood distribution of a large set of drugs from structurally derived descriptors using partial least-squares (PLS) modeling. *Journal of Chemical Information and Computer Sciences*, 39, 396–404.
- MACCS. (2011). MACCS structural keys. San Diego, CA: Accelrys.
- Mannhold, R., & Dross, K. (1996). Calculation procedures for molecular lipophilicity: A comparative study. *Quantitative Structure-Activity Relationships*, 15, 403–409.
- Mannhold, R., Poda, G. I., Ostermann, C., & Tetko, I. V. (2009). Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds. *Journal of Pharmaceutical Sciences*, 98, 861–893.

- Martin, Y. C. (2005). A bioavailability score. *Journal of Medicinal Chemistry* 48, 3164–3170.
- Martin, Y. C. (1978). *Quantitative Drug Design*. New York: A Critical Introduction. Marcel Dekker.
- Meanwell, N. A. (2016). Improving Drug Design: An Update on Recent Applications of Efficiency Metrics, Strategies for Replacing Problematic Elements, and Compounds in Nontraditional Drug Space. *Chemical Research in Toxicology*, 29, 564–616.
- Mekenyan, O., & Bonchev, D. (1986). Oasis method for predicting biological-activity of chemical-compounds. *Acta Pharmaceutica Jugoslavica*, 36, 225–237.
- Mitchell, M., (1998). An introduction to genetic algorithms. MIT press.
- Moda, T. L., Montanari, C. A., & Andricopulo, A. D. (2007). Hologram QSAR model for the prediction of human oral bioavailability. *Bioorganic & Medicinal Chemistry*, 15, 7738–7745.
- MOE. (2016). Molecular operating environment (MOE), 2013.08; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7.
- Moura Barbosa, A. J., & Del Rio, A. (2012). Freely accessible databases of commercial compounds for high-throughput virtual screenings. *Current Topics in Medicinal Chemistry*, 12, 866–877.
- Mostrag-Szlichtyng, A., & Worth, A. (2010). *Review of QSAR models and software tools for predicting biokinetic properties*. Comm: Luxemb. Eur.
- Muir, R. M., Fujita, T., & Hansch, C. (1967). Structure-activity relationship in the auxin activity of mono-substituted phenylacetic acids. *Plant Physiology*, 42, 1519–1526.
- Mysinger, M. M., Carchia, M., Irwin, J. J., & Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55, 6582–6594. doi:10.1021/jm300687e.
- Narayanan, R., & Gunturi, S. B. (2005). In silico ADME modelling: Prediction models for blood-brain barrier permeation using a systematic variable selection method. *Bioorganic & Medicinal Chemistry*, 13, 3017–3028.
- Navratilova, I., Myszka, D. G., & Rich, R. L. (2007). Probing membrane protein interactions with real-time biosensor technology. *Biophysical Analysis of Membrane Proteins: Investigating Structure and Function*, 121–140.
- Netzeva, T. I., Worth, A. P., Aldenberg, T., Benigni, R., Cronin, M. T., Gramatica, P., et al. (2005). Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. *ATLA* 33, 155–173.
- Nicolotti, O., Gillet, V. J., Fleming, P. J., & Green, D. V. (2002). Multiobjective optimization in quantitative structure-activity relationships: Deriving accurate and interpretable QSARs. *Journal of Medicinal Chemistry*, 45, 5069–5080.
- Oprea, T. I. (2000). Property distribution of drug-related chemical databases. *Journal of Computer-Aided Molecular Design*, 14, 251–264.
- Owens, P. K., Raddad, E., Miller, J. W., Stille, J. R., Olovich, K. G., Smith, N. V., et al. (2015). A decade of innovation in pharmaceutical R&D: The chorus model. *Nature Reviews Drug Discovery*, 14, 17–28.
- Pajouhesh, H., & Lenz, G. R. (2005). Medicinal chemical properties of successful central nervous system drugs. *NeuroRx*, 2, 541–553.
- Papadatos, G., Gaulton, A., Hersey, A., & Overington, J. P. (2015). Activity, assay and target data curation and quality in the ChEMBL database. *Journal of Computer-Aided Molecular Design*, 29, 885–896.
- Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, et al. (2010). ArrayExpress update—An archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Research*, gkq1040.
- Paterlini, S., & Minerva, T. (2010). Regression model selection using genetic algorithms. In *Proceedings of the 11th WSEAS International Conference on Neural Networks and 11th WSEAS International Conference on Evolutionary Computing and 11th WSEAS International Conference on Fuzzy Systems*, pp. 19–27.

- Pastor, M., Cruciani, G., Mclay, I., Pickett, S., & Clementi, S. (2000). Grid-independent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. *Journal of Medicinal Chemistry*, *43*, 3233–3243.
- Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., et al. (2010). How to improve R&D productivity: The pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, *9*, 203–214.
- Pham-The, H., González-Álvarez, I., Bermejo, M., Garrigues, T., Le-Thi-Thu, H., & Cabrera-Pérez, M. Á. (2013). The use of rule-based and QSPR approaches in ADME profiling: A case study on caco-2 permeability. *Molecular Informatics*, *32*, 459–479.
- Platts, J. A., Abraham, M. H., Zhao, Y. H., Hersey, A., Ijaz, L., & Butina, D. (2001). Correlation and prediction of a large blood–brain distribution data set—an LFER study. *European Journal of Medicinal Chemistry*, *36*, 719–730.
- Pliška, V., Testa, B., & van de Waterbeemd, H. (1996). Lipophilicity: The empirical tool and the fundamental objective, an introduction. In V. Pliška, B. Testa, P. -D. H. van de Waterbeemd (Eds.), *Lipophilicity in drug action and toxicology* (pp. 1–6). Wiley-VCH Verlag GmbH.
- Polanski, J. (2009). Receptor dependent multidimensional QSAR for modeling drug-receptor interactions. *Current Medicinal Chemistry*, *16*, 3243–3257.
- Puzyn, T., Leszczynski, J., & Cronin, M. T. (2010). *Recent advances in QSAR studies: Methods and applications*. Springer Science & Business Media.
- Qiu, T., Qiu, J., Feng, J., Wu, D., Yang, Y., Tang, K., et al. (2016). The recent progress in proteochemometric modelling: Focusing on target descriptors, cross-term descriptors and application scope. *Briefings in Bioinformatics*. bbw004.
- Rekker, R. F., & Mannhold, R. (1992). *Calculation of drug lipophilicity: The hydrophobic fragmental constant approach*. Wiley-VCH.
- Rich, R. L., & Myszka, D. G. (2000). Advances in surface plasmon resonance biosensor analysis. *Current Opinion in Biotechnology*, *11*, 54–61.
- Rodgers, S. L., Davis, A. M., Tomkinson, N. P., & van de Waterbeemd, H. (2011). Predictivity of simulated ADME AutoQSAR models over time. *Molecular Informatics*, *30*, 256–266.
- Rodgers, S. L., Davis, A. M., & van de Waterbeemd, H. (2007). Time-series QSAR analysis of human plasma protein binding data. *QSAR & Combinatorial Science*, *26*, 511–521.
- Rogge, M. C., & Taft, D. R. (Eds.). (2010) preclinical drug development second edition. In *Drugs and the pharmaceutical sciences* (Vol.187). CRS Press, Taylor and Francis Group.
- Rowley, M., Kulagowski, J. J., Watt, A. P., Rathbone, D., Stevenson, G. I., Carling, R. W., et al. (1997). Effect of plasma protein binding on in vivo activity and brain penetration of glycine/NMDA receptor antagonists. *Journal of Medicinal Chemistry*, *40*, 4053–4068. doi:10.1021/jm970417o.
- Roy, K., Mitra, I., Kar, S., Ojha, P. K., Das, R. N., & Kabir, H. (2012). Comparative studies on some metrics for external validation of QSPR models. *Journal of Chemical Information and Modeling*, *52*, 396–408. doi:10.1021/ci200520g.
- Roy, P. P., Paul, S., Mitra, I., & Roy, K. (2009). On two novel parameters for validation of predictive QSAR models. *Molecules*, *14*, 1660–1701.
- Rücker, C., Rücker, G., & Meringer, M. (2007). y-Randomization and its variants in QSPR/QSAR. *Journal of Chemical Information and Modeling*, *47*, 2345–2357.
- Rutenber, E. E., & Stroud, R. M. (1996). Binding of the anticancer drug ZD1694 to E. coli thymidylate synthase: Assessing specificity and affinity. *Structure*, *4*, 1317–1324.
- Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., & Todeschini, R. (2012). Comparison of different approaches to define the applicability domain of QSAR models. *Molecules*, *17*, 4791–4810.
- Saiakhov, R. D., Stefan, L. R., & Klopman, G. (2000). Multiple computer-automated structure evaluation model of the plasma protein binding affinity of diverse drugs. *Perspectives in Drug Discovery and Design*, *19*, 133–155.
- Sakiyama, Y. (2009). The use of machine learning and nonlinear statistical tools for ADME prediction. *Expert Opinion on Drug Metabolism & Toxicology*, *5*, 149–169.

- Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R., & Sherman, W. (2013). Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, *27*, 221–234.
- Satyanarayanan, S.D. (2011). *Drug design and discovery: Methods and protocols*. Humana Press.
- Schindler, T., Bornmann, W., Pellicena, P., Miller, W. T., Clarkson, B., & Kuriyan, J. (2000). Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. *Science*, *289*, 1938–1942.
- Schneider, G. (2010). Virtual screening: An endless staircase? *Nature Reviews Drug Discovery*, *9*, 273–276.
- Sedykh, A., Zhu, H., Tang, H., Zhang, L., Richard, A., Rusyn, I., et al. (2011). Use of in vitro HTS-derived concentration-response data as biological descriptors improves the accuracy of QSAR models of in vivo toxicity. *Environmental Health Perspectives*, *119*, 364.
- Sherer, E. C., Verras, A., Madeira, M., Hagemann, W. K., Sheridan, R. P., Roberts, D., et al. (2012). QSAR Prediction of passive permeability in the LLC-PK1 cell line: Trends in molecular properties and cross-prediction of caco-2 permeabilities. *Molecular Informatics*, *31*, 231–245.
- Sheridan, R. P. (2014). Global quantitative structure-activity relationship models vs selected local models as predictors of off-target activities for project compounds. *Journal of Chemical Information and Modeling*, *54*, 1083–1092.
- Sheridan, R. P., Korzekwa, K. R., Torres, R. A., & Walker, M. J. (2007). Empirical regioselectivity models for human cytochromes P450 3A4, 2D6, and 2C9. *Journal of Medicinal Chemistry*, *50*, 3173–3184.
- Sheridan, R. P., McMasters, D. R., Voigt, J. H., & Wildey, M. J. (2015). eCounterscreening: Using QSAR predictions to prioritize testing for off-target activities and setting the balance between benefit and risk. *Journal of Chemical Information and Modeling*, *55*, 231–238.
- Sohn, Y. S., Park, C., Lee, Y., Kim, S., Thangapandian, S., Kim, Y., et al. (2013). Multi-conformation dynamic pharmacophore modeling of the peroxisome proliferator-activated receptor γ for the discovery of novel agonists. *Journal of Molecular Graphics and Modelling*, *46*, 1–9.
- Speck-Planche, A., & Cordeiro, M. N. D. S. (2015). Multitasking models for quantitative structure—Biological effect relationships: Current status and future perspectives to speed up drug discovery. *Expert Opinion on Drug Discovery*, *10*, 245–256.
- Spowage, B. M., Bruce, C. L., & Hirst, J. D. (2009). Interpretable correlation descriptors for quantitative structure—Activity relationships. *Journal of Cheminformatics*, *1*, 22.
- Stegmaier, K., Ross, K. N., Colavito, S. A., O'Malley, S., Stockwell, B. R., & Golub, T. R. (2004). Gene expression-based high-throughput screening (GE-HTS) and application to leukemia differentiation. *Nature Genetics*, *36*, 257–263.
- Stuper, A. J., & Jurs, P. C. (1976). ADAPT: A computer system for automated data analysis using pattern recognition techniques. *Journal of Chemical Information and Modeling*, *16*, 99–105. doi:10.1021/ci60006a014.
- Suenderhauf, C., Hammann, F., & Huwyler, J. (2012). Computational prediction of blood–brain barrier permeability using decision tree induction. *Molecules*, *17*, 10429–10445.
- Sun, H. (2004). A universal molecular descriptor system for prediction of logP, logS, logBB, and absorption. *Journal of Chemical Information and Computer Sciences*, *44*, 748–757.
- Swift, R. V., & Amaro, R. E. (2013). Back to the future: Can physical models of passive membrane permeability help reduce drug candidate attrition and move us beyond QSPR? *Chemical Biology & Drug Design*, *81*, 61–71.
- Talafous, J., Sayre, L. M., Miesal, J. J., & Klopman, G. (1994). META. 2. A dictionary model of mammalian xenobiotic metabolism. *Journal of Chemical Information and Computer Sciences*, *34*, 1326–1333.
- Tao, L., Zhang, P., Qin, C., Chen, S. Y., Zhang, C., Chen, Z., et al. (2015). Recent progresses in the exploration of machine learning methods as in-silico ADME prediction tools. *Advanced Drug Delivery Reviews*, *86*, 83–100.

- Tarcsay, Á., Nyíri, K., & Keserű, G. M. (2012). Impact of lipophilic efficiency on compound quality. *Journal of Medicinal Chemistry*, 55, 1252–1260.
- Terfloth, L., Bienfait, B., & Gasteiger, J. (2007). Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. *Journal of Chemical Information and Modeling*, 47, 1688–1701.
- Testa, B., Balmat, A.-L., & Long, A. (2004). Predicting drug metabolism: Concepts and challenges. *Pure and Applied Chemistry*, 76, 907–914.
- Testa, B., Balmat, A.-L., Long, A., & Judson, P. (2005a). Predicting drug metabolism—An evaluation of the expert system METEOR. *Chemistry & Biodiversity*, 2, 872–885.
- Testa, B., Vistoli, G., & Pedretti, A. (2005b). Musings on ADME predictions and structure-activity relations. *Chemistry & Biodiversity*, 2, 1411–1427. doi:10.1002/cbdv.200590115.
- Testa, B. (2009). Drug metabolism for the perplexed medicinal chemist. *Chemistry & Biodiversity*, 6, 2055–2070.
- Tetko, I. V., Poda, G. I., Ostermann, C., & Mannhold, R. (2009). Large-scale evaluation of log P predictors: Local corrections may compensate insufficient accuracy and need of experimentally testing every other compound. *Chemistry & Biodiversity*, 6, 1837–1844.
- Tetko, I. V., Tanchuk, V. Y., & Villa, A. E. (2001). Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. *Journal of Chemical Information and Computer Sciences*, 41, 1407–1421.
- Thiel-Demby, V. E., Humphreys, J. E., St. John Williams, L. A., Ellens, H. M., Shah, N., Ayrton, et al. (2008). Biopharmaceutics classification system: Validation and learnings of an in vitro permeability assay. *Molecular Pharmaceutics* 6, 11–18.
- Tian, S., Li, Y., Wang, J., Zhang, J., & Hou, T. (2011). ADME evaluation in drug discovery. 9. Prediction of oral bioavailability in humans based on molecular properties and structural fingerprints. *Molecular Pharmaceutics*, 8, 841–851.
- Tilley, J. W., Chen, L., Fry, D. C., Emerson, S.D., Powers, G.D., Biondi, D., et al. (1997). Identification of a small molecule inhibitor of the IL-2/IL-2R α receptor interaction which binds to IL-2. *Journal of the American Chemical Society* 119, 7589–7590.
- Todeschini, R., & Consonni, V. (2009). *Molecular descriptors for chemoinformatics*, Volume 41 (2 Volume Set). Wiley.
- Tropsha, A., Gramatica, P., & Gombar, V. K. (2003). The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science*, 22, 69–77. doi:10.1002/qsar.200390007.
- Tsantili-Kakoulidou, A., & Agrafiotis, D. K. (2011). The 18th european symposium on quantitative structure-activity relationships. *Expert Opinion on Drug Discovery*, 6, 453–456.
- Tsopelas, F., Vallianatou, T., & Tsantili-Kakoulidou, A. (2016a). The potential of immobilized artificial membrane chromatography to predict human oral absorption. *European Journal of Pharmaceutical Sciences*, 81, 82–93.
- Tsopelas, F., Vallianatou, T., & Tsantili-Kakoulidou, A. (2016b). Advances in immobilized artificial membrane (IAM) chromatography for novel drug discovery. *Expert Opinion on Drug Discovery*, 11, 473–488.
- Ursu, O., Rayan, A., Goldblum, A., & Oprea, T. I. (2011). Understanding drug-likeness. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1, 760–781.
- Usansky, H. H., & Sinko, P. J. (2005). Estimating human drug oral absorption kinetics from Caco-2 permeability using an absorption-disposition model: Model development and evaluation and derivation of analytical solutions for k_a and F_a . *Journal of Pharmacology and Experimental Therapeutics*, 314, 391–399.
- Vallianatou, T., Lambrinidis, G., Giaginis, C., Mikros, E., & Tsantili-Kakoulidou, A. (2013). Analysis of PPAR- α/γ Activity by Combining 2-D QSAR and Molecular Simulation. *Molecular Informatics*, 32, 431–445.
- van de Waterbeemd, H., Camenisch, G., Folkers, G., & Raevsky, O. A. (1996). Estimation of Caco-2 cell permeability using calculated molecular descriptors. *Quantitative Structure-Activity Relationships*, 15, 480–490.

- van de Waterbeemd, H., & Smith, D. A., (2001). Relations of molecular properties with drug disposition: The cases of gastrointestinal absorption. *Pharmacokinetic Optimization in Drug Research: Biological, Physicochemical, and Computational Strategies*, 51.
- Van de Waterbeemd, H., & Testa, B. (1987). The parametrization of lipophilicity and other structural properties in drug design. *Advances in Drug Research*, 16, 85–225.
- Varghese, J. N. (1999). Development of neuraminidase inhibitors as anti-influenza virus drugs. *Drug Development Research*, 46, 176–196.
- Vastag, M., & Keseru, G. M. (2009). Current in vitro and in silico models of blood–brain barrier penetration: a practical view. *Current Opinion in Drug Discovery & Development*, 12, 115–124.
- Vasudevan, S. R., & Churchill, G. C. (2009). Mining free compound databases to identify candidates selected by virtual screening. *Expert Opinion on Drug Discovery*, 4, 901–906.
- Veber, D. F., Johnson, S. R., Cheng, H.-Y., Smith, B. R., Ward, K. W., & Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, 45, 2615–2623.
- Vedani, A., Briem, H., Dobler, M., Dollinger, H., & McMasters, D. R. (2000). Multiple-conformation and protonation-state representation in 4D-QSAR: the neurokinin-1 receptor system. *Journal of Medicinal Chemistry*, 43, 4416–4427.
- Vedani, A., Dobler, M., & Lill, M. A. (2005). Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. *Journal of Medicinal Chemistry*, 48, 3700–3703.
- Veerasingh, R., Rajak, H., Jain, A., Sivadasan, S., Varghese, C. P., & Agrawal, R. K. (2011). Validation of QSAR models-strategies and importance. *International Journal of Drug Design & Discovery*, 3, 511–519.
- Vilar, S., Chakrabarti, M., & Costanzi, S. (2010). Prediction of passive blood–brain partitioning: straightforward and effective classification models based on in silico derived physicochemical descriptors. *Journal of Molecular Graphics and Modelling*, 28, 899–903.
- Volpe, D. A. (2008). Variability in Caco-2 and MDCK cell-based intestinal permeability assays. *Journal of Pharmaceutical Sciences*, 97, 712–725.
- Votano, J. R., Parham, M., Hall, L. M., Hall, L. H., Kier, L. B., Oloff, S., et al. (2006). QSAR modeling of human serum protein binding with several modeling techniques utilizing structure-information representation. *Journal of Medicinal Chemistry*, 49, 7169–7181. doi:10.1021/jm051245v.
- Wager, T. T., Villalobos, A., Verhoest, P. R., Hou, X., & Shaffer, C. L. (2011). Strategies to optimize the brain availability of central nervous system drug candidates. *Expert Opinion on Drug Discovery*, 6, 371–381.
- Wang, N.-N., Dong, J., Deng, Y.-H., Zhu, M.-F., Wen, M., Yao, Z.-J., et al. (2016). ADME properties evaluation in drug discovery: Prediction of caco-2 cell permeability using a combination of NSGA-II and boosting. *Journal of Chemical Information and Modeling*, 56, 763–773.
- Wang, Y., Xiao, J., Suzek, T.O., Zhang, J., Wang, J., Zhou, Z., et al. (2012). PubChem's BioAssay database. *Nucleic Acids Research* 40, D400–D412.
- Wessel, M. D., Jurs, P. C., Tolani, J. W., & Muskal, S. M. (1998). Prediction of human intestinal absorption of drug compounds from molecular structure. *Journal of Chemical Information and Computer Sciences*, 38, 726–735.
- Wiesmann, C., Christinger, H. W., Cochran, A. G., Cunningham, B. C., Fairbrother, W. J., Keenan, C. J., et al. (1998). Crystal structure of the complex between VEGF and a receptor-blocking peptide. *Biochemistry (Mosc)*, 37, 17765–17772.
- Willett, P. (2004). Evaluation of molecular similarity and molecular diversity methods using biological activity data in methods in molecular biology. In J. Bajorath (Ed.), *Cheminformatics: Concepts, methods, and tools for drug discovery* (Vol. 275). Totowa, NJ: Humana Press Inc.
- Williams, G. (2012). A searchable cross-platform gene expression database reveals connections between drug treatments and disease. *BMC Genomics*, 13, 1.

- Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58, 109–130.
- Wong, W. W., & Burkowski, F. J. (2009). A constructive approach for discovering new drug leads: Using a kernel methodology for the inverse-QSAR problem. *Journal of Cheminformatics*, 1, 1.
- Worth, A. P., Hartung, T., & Van Leeuwen, C. J. (2004). The role of the European centre for the validation of alternative methods (ECVAM) in the validation of (Q) SARs. *SAR and QSAR in Environmental Research*, 15, 345–358.
- Yee, S. (1997). In vitro permeability across caco-2 cells (colonic) can predict in vivo (small intestinal) absorption in man—Fact or myth. *Pharmaceutical Research*, 14, 763–766.
- Yera, E. R., Cleves, A. E., & Jain, A. N. (2014). Prediction of off-target drug effects through data fusion. In *Pacific Symposium on Biocomputing*. NIH Public Access, p. 160.
- Yusof I., & Segall, M. D. (2013). Considering the impact drug-like properties have on the chance of success. *Drug Discovery Today* 18, 659–66.
- Zhao, P., Rowland, M., & Huang, S.-M. (2012). Best practice in the use of physiologically based pharmacokinetic modeling and simulation to address clinical pharmacology regulatory questions. *Clinical Pharmacology and Therapeutics*, 92, 17–20.
- Zhao, Y. H., Le, J., Abraham, M. H., Hersey, A., Eddershaw, P. J., Luscombe, C. N., Boutina, D., et al. (2001). Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure-activity relationship (QSAR) with the Abraham descriptors. *Journal of Pharmaceutical Sciences* 90, 749–784.
- Zsila, F. (2013). Subdomain IB is the third major drug binding region of human serum albumin: Toward the three-sites model. *Molecular Pharmaceutics*, 10, 1668–1682.

Strategy for Identification of Nanomaterials' Critical Properties Linked to Biological Impacts: Interlinking of Experimental and Computational Approaches

Iseult Lynch, Antreas Afantitis, Georgios Leonis, Georgia Melagraki and Eugenia Valsami-Jones

Abstract Significant progress has been made over the last 10 years towards understanding those characteristics of nanoscale particles which correlate with enhanced biological activity and/or toxicity, as the basis for development of predictive tools for risk assessment and safer-by-design strategies. However, there are still a number of disconnects in the nanosafety workflow that hamper rapid progress towards full understanding of nano-specific mechanisms of action and nanomaterials (NMs)-induced adverse outcome pathways. One such disconnect is between physico-chemical characteristics determined experimentally as part of routine NMs characterisation, and the ability to predict a NM's uptake and impacts on biological systems based on its pristine physico-chemical characteristics. Identification of critical properties (physico-chemical descriptors) that confer the ability to induce harm in biological systems *under the relevant exposure conditions* is central, in order to enable both prediction of impacts from related NMs [via quantitative property-activity or structure-activity relationships (QPARs/QSARs)] and

I. Lynch (✉) · E. Valsami-Jones

School of Geography, Earth and Environmental Sciences University of Birmingham,
Birmingham B15 2TT, UK
e-mail: i.lynch@bham.ac.uk

E. Valsami-Jones

e-mail: E.ValsamiJones@bham.ac.uk

A. Afantitis · G. Leonis · G. Melagraki (✉)

Novamechanics Ltd., Nicosia, Cyprus

e-mail: melagraki@novamechanics.com; melagraki@insilicolab.eu

A. Afantitis

e-mail: afantitis@novamechanics.com

G. Leonis

e-mail: leonis@novamechanics.com

A. Afantitis · G. Melagraki

InSilicoLab L.P., Athens, Greece

© Springer International Publishing AG 2017

K. Roy (ed.), *Advances in QSAR Modeling*, Challenges and Advances

in Computational Chemistry and Physics 24, DOI 10.1007/978-3-319-56850-8_10

development of strategies to ensure that these features are avoided in NM production in the future (“safety by design”). For this purpose, we have launched the Enalos InSilico platform, which is dedicated to the dissemination of our developed in silico workflows for NM risk assessment. So far, two predictive models have been made available online. The first tool is a Quantitative Nanostructure-Activity Relationship (QNAR) model for the prediction of the cellular uptake of NMs in pancreatic cancer cells and the second is an online tool for in silico screening of iron oxide NMs with a predictive classification model for their toxicological assessment.

Keywords Nanomaterial characterization • Physico-chemical/structural properties • Data mining • Machine learning for nanomaterial data • Quantitative Nanostructure-Activity Relationship (QNAR)

List of Abbreviations

AOP	Adverse Outcome Pathway
BSAI	Biological Surface Adsorption Index
CCC	Critical Coagulation Concentration
EU	European Union
HOMO	Highest Occupied Molecular Orbital
LUMO	Lowest Unoccupied Molecular Orbital
MIE	Molecular Initiating Event
NM	Nanomaterial
NP	Nanoparticle
PS-NH ₂	Amino-functionalized polystyrene
PZC	Point of Zero Charge
QNAR	Quantitative Nanomaterial-Activity Relationship
QPAR	Quantitative Property-Activity Relationship
QSAR	Quantitative Structure-Activity Relationship
ROS	Reactive Oxygen Species
TEM	Transmission Electron Microscopy

1 Introduction

Nanomaterials (NMs) are a highly diverse group of chemicals, defined mainly by their small size, which ranges from 1 to 100 nm, but varying enormously regarding their physico-chemical properties, such as composition, shape, surface charge, crystallinity, and reactivity, among others (Stamm 2011). Researchers in the field of NM science are struggling to associate the primary properties of NMs with their biological reactivity and toxicity (Valsami-Jones 2015; Nel 2015), as well as formulating appropriate methodologies to understand and utilize them to their full potential. Due to the widespread application and commercial usefulness of NMs in

products ranging from industrial to consumer goods (Tsuzuki 2009), it is necessary to view NMs from a regulatory perspective. However, this is particularly challenging, in part because methods to identify many of the important physico-chemical properties are lacking, or are not yet sufficiently validated (von der Kammer et al. 2012). Consequently, there is an extensive literature indicating that only size measurements are currently reliable to establish a regulatory definition of NMs (Linsinger 2012). Three main NM-associated concerns have been implicated in making regulation, read-across and impact prediction of NMs problematic: (1) the fact that many properties are non-scalable, (2) the need to distinguish between intrinsic versus extrinsic (i.e., context dependent) properties, (Lynch et al. 2014a) and (3) the fact that many properties are interlinked (e.g., changing one property may induce changes to another) which renders the description of property-activity relationships arduous and makes the development of systemic libraries of NMs challenging. These concerns are presented below in detail.

As part of continuing EU efforts to define NMs for regulatory purposes, the Joint Research Centre (JRC) of the European Commission have classified physico-chemical parameters of NMs into those that scale with size (scalable) and those that display unique nanoscale characteristics below a certain size (non-scalable) (Lövestam 2010). Examples of non-scalable properties include confinement effects, such as the broad HOMO-LUMO gap (or the related band gap) of semiconductor NMs that increases drastically for diameters below 5 nm, thermal properties, such as the exponential decrease of the melting point of In and Sn NMs below a diameter of 15 nm, and the solubility via dependence on the surface tension, which deviates significantly from the classical behaviour when particle sizes drop below 25 nm (Lövestam 2010). Size-dependent crystallinity also alters the interface properties of NMs, such as surface reaction rates, adsorption capacity, catalytic processes and redox potential, which control molecular processes that are related to diverse cell functions. The challenge is evident since non-scalable properties are material-dependent, and there is no straightforward, material-independent relation between particle size and properties or functions (Lövestam 2010).

In connection with the non-scalable parameters, several NM properties depend on the context in which they are studied, meaning that they are affected by the "environment". For example, the layer of biological molecules that surrounds certain NMs upon dispersion in a biological fluid has been implicated in conferring a "biological identity" (Walczyk 2010), which derives from the elemental synthetic identity, (Fadeel 2013) as this determines which biomolecules bind to the surface (e.g., through electrostatic and hydrophobic interactions, as well as favourable entropy contributions) (Dawson 2007). A critical article from Yang et al. outlines the relationship among NM surface properties, the properties of the surrounding medium (e.g., pH, ionic strength, salt composition) and the properties of the bio- or macro-molecules under study (Yang 2013). A related approach has been considered to describe and predict NM interactions in the natural environment, thus highlighting the ecological identity of NMs (Lynch et al. 2014a). A suggested classification framework assumed that NM toxicity can be predicted as the sum of three

“quantifiable” parameters (principal components). These parameters include the diversity of modes of action of NMs, and are classified as *intrinsic* properties (e.g., structural features), *extrinsic* properties (e.g., surface interactions, changes induced upon binding of biomolecules and other environmental interactions), and composition (Lynch et al. 2014b). Several NM physico-chemical properties belong to the intrinsic category (for example, strain which also includes shape, porosity, structure and HOMO-LUMO gap), and our consideration allows the construction of a scale depicting their relative contributions to this category. Similarly, conformational changes due to binding of biomolecules, such as protein unfolding, receptor activation, membrane damage, and fibrillation comprise the second category. Chemical composition is the third category and properties linked to the inherent molecular toxicity, charge, hydrophobicity and coating (also associated with both the intrinsic and extrinsic descriptions) are important (Lynch et al. 2014a). Table 1 provides an initial estimation of the key physico-chemical features that are considered crucial for NM toxicity, and whether they are likely to be context-dependent and thus require additional characterisation under the relevant exposure conditions as well as in the pristine form. The importance of such changes to the NMs properties and their adsorbed layers of biomolecules (the so-called biological and ecological identities of NMs), in terms of predicting NM uptake and toxicity is one of the key questions to be addressed by QSARs/QPARs currently. As yet however, limited attention has been paid to the physical transformations that NMs themselves undergo during, for example, environmental ageing (Lowry 2012) in terms of incorporation into QSARs.

The third challenge is the inter-dependency of many NM properties. Unfortunately, the exact relationships governing these interdependencies have yet to be established, in part due to the lack of available libraries of systematically varied individual NM properties. This would require property variation in a precise manner to identify crucial parameters driving toxicity, and to evaluate toxicity thresholds for various descriptors. However, the development of systematically varied NMs libraries is hampered by the fact that the variation of one property may inadvertently induce changes to several others, for example synthesis strategies to change the shape or length of a NM may require use of different templating molecules that result in differences in the surface chemistry of the particles also (Soler 2007; Bussy 2012; Zhang 2012b). This is illustrated in Fig. 1 for two different NMs (Au and ZnO, being typical examples of a metal and a soluble metal oxide, respectively), where the obviously interlinked properties are highlighted, although other inter-linkages likely also exist. This interdependence of physico-chemical properties also contributes to the inadequacy of the scientific efforts to date to obtain a set of agreed descriptors for NMs classification [even on a limited set of physico-chemical end-points against which to characterise NMs (see Stefaniak 2013 and discussion therein)].

The schematic representation (Fig. 1) shows some of the descriptor space that can, in principle, be varied through development of NM libraries and also demonstrates that (i) not all descriptors are relevant to all NM types, and (ii) not all parameters can be varied independently as in some cases the method used to vary

Table 1 Initial assessment of the potential context-dependent changes in physico-chemical properties of NMs

Parameter/Descriptor	Context dependent?	Potential impacts of surroundings
Size/Size distribution	Yes	In the environment, most likely decreased by binding of natural organic matter (stabilization). Protein binding may lead to either increased or decreased size via bridging or steric stabilisation pH/ionic strength may alter agglomeration
Surface area	Yes	Aggregation/agglomeration will reduce available surface area
Purity (particle/dispersant)	Maybe	Impurities/dispersants may be more effectively released from NM surface under different environmental conditions
Dissolution potential	Yes	pH, ionic strength, redox potential and adsorbed biomolecules affect dissolution rate
Photochemical activity	Most likely	Differences in pH and ionic strength and presence/absence of organic matter may affect electron transfer and result in protonation of different excited states
Surface charge/chemistry	Yes	Binding of ions/biomolecules may confer a different charge/charge distribution and surface groups but this may be dynamic
Hydrophobicity	Yes	Binding of biomolecules typically results in a more hydrophilic surface presentation, although may be dynamic
Redox activity	Most likely	Different surfaces/coatings/bound ligands may result in different radical species being generated (Li 2013)
Shape	Most likely	Agglomeration will result in different overall shape. Bundling/unbundling of nanotubes is an example
Crystal structure	Unlikely	Structure is a bulk property, established during the formation of an NM and cannot change by processes occurring on the surface, unless if the NM dissolves completely and re-precipitates
Porosity/Surface defects	Most likely	Though dependent on pore size or nature of defect, most likely decreased due to biomolecule absorption; may also be influenced by dissolution if NMs do not dissolve congruently or are a mixed phase

one parameter also changes another. The examples show gold (Au) NMs and ZnO NMs, which are used to represent easily versus poorly soluble NMs. In the example of Au NMs, several of the parameters are automatically ruled out as the atomic structure and subsequent packing of Au molecules does not result in particles with different crystal structures [at least at sizes beyond ~ 1 nm whereupon the cluster structure is templated into the particles (Wells 2015)]. In the example of ZnO NMs, it is clear that changing the capping agent will impact on dissolution potential, and surface properties such as charge and hydrophobicity, and thus the capping agent cannot be varied in insolation from other (interlinked) properties.

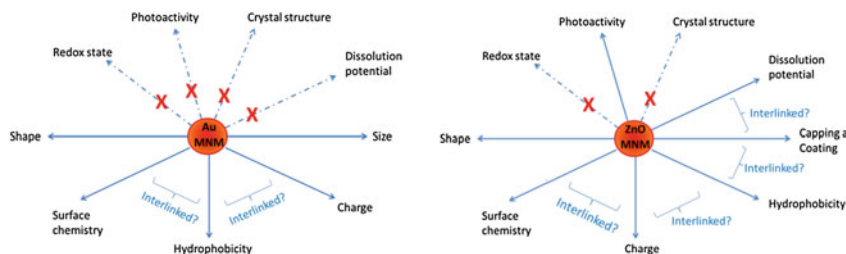


Fig. 1 Graphical presentation of the principles of NM library development and how systematic variation of one parameter can result in changes to other (interlinked) parameters

In addition to the direct interdependencies of physico-chemical properties, there is also the potential for additive (cooperative) or competitive effects in terms of how NMs properties result in interactions with, and impacts on, living systems. Enhanced binding to target cells using multiple physical and chemical interactions, and a range of distances where the addition of two effective repulsive interactions became an attraction, has been demonstrated (Nap 2013). Such models give insight into the competing and highly non-additive nature of different effective interactions in nanoscale systems in constrained environments, such as are ubiquitous in synthetic and biological systems, and suggest that these should be taken into account in the development of QNARs.

Besides the aforementioned NM challenges, there are also practical concerns among the diverse communities working with NMs to communicate with each other, regarding terminology, limitations of synthetic procedures, characterisation and modelling approaches, etc. as well realistic plans to produce the experimental results needed to underpin and validate models in the short and medium terms. Thus, there are two major obstacles regarding successful development of QNARs: (a) the lack of adequate and systematic experimental data (which requires high-quality systematically varied NM libraries) and (b) the currently limited knowledge on mechanisms of toxic action of NMs under realistic exposure and ageing conditions. Subsequent sections of this chapter offer insights into the current state of the art in terms of NM libraries and the QSARs that have been developed using these libraries, as well as the present understanding of the main physico-chemical properties of toxicological relevance (see also Table 1).

An understanding of the relationship between the physico-chemical properties of a particular NM and its *in vitro* and *in vivo* behaviour would provide the basic information for assessing toxic response and more importantly may yield predictive models for sub-classes of NMs allowing grouping of NMs in a similar manner to that applied for chemicals. Thus, we also outline current efforts to bridge the current disconnect between the modelling and experimental communities, and thereby to enhance progress. Among the successes listed so far are the development of a tool for surface chemistry, which is challenging experimentalists to re-think the way NMs and their surfaces are described and connect these with cellular uptake, and a

model for toxicity assessment of iron oxides with different core, coating and surface modifications.

Here, we propose a living classification system that incorporates the context-dependent evolution of several physico-chemical features of NMs (as shown in Table 1) and evolves continuously with new data generation and as new patterns emerge. Practically, this process may be likened to a modelling scheme, where the initial steps constitute a training process, populating it with data, which associate physico-chemical descriptors of NMs (including aged and biologically or environmentally transformed forms) with biological impacts. This would be the cornerstone for a second phase of testing the classification system against less well defined NMs to ensure that the predicted classifications match the data generated, and then following an approach where less and less experimental data is needed in order to evaluate classification and safety implications. This will produce a valuable set of tools for QSARs/QPARs and prediction of NM effects for risk assessment, regulation, and re-design of NMs synthesis according to the principles of safety by design to minimise the occurrence and effect of descriptors found to link directly to a toxicological mechanism or end-point as shown in Fig. 2.

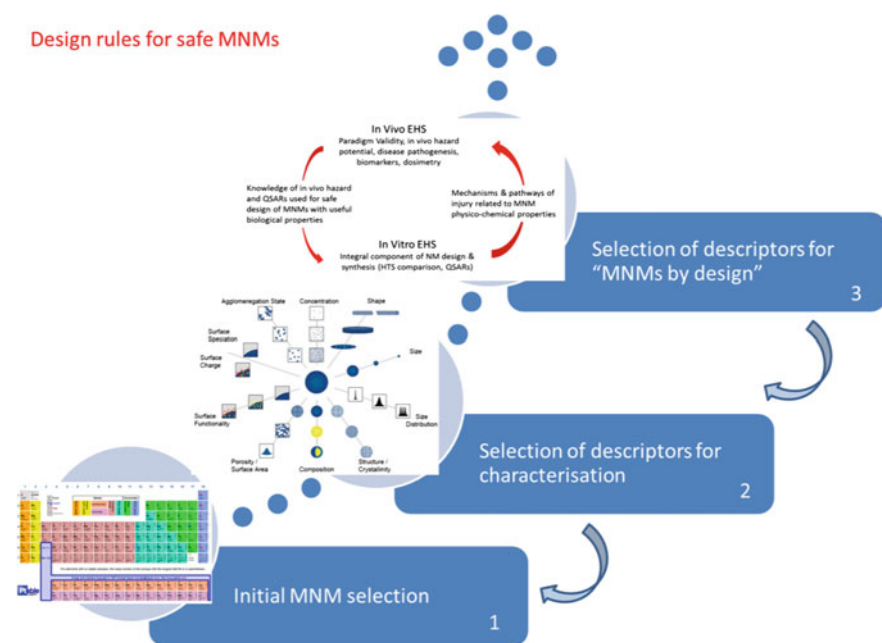


Fig. 2 Spiralling-type approach to NM classification, where earlier phases are “inspired” by tentative efforts at defining the later stages. *Note* graphical illustrations for stages 2 and 3 are from Hassellöv (2009) and Meng (2009), respectively

2 NM Physico-chemical Properties of Toxicological Relevance

The establishment of a safe nanotechnology necessitates the development of evaluation procedures to determine hazardous NM properties that could be modified to improve NM safety (George 2012).

- (i) the release of toxic compounds from NMs (e.g., Cd from quantum dots)—i.e., **NM dissolution**;
- (ii) the direct effects induced after physical contact with NMs, influenced by their size, shape and surface features, and which interfere with important biological functions—i.e., **NM interactions**; (so-called *extrinsic* factors)
- (iii) the inherent features of the NM, such as photochemical and redox properties resulting from band gap or crystalline phase—i.e., **NM (surface) specific effects**; (so-called *intrinsic* factors), and
- (iv) the capacity of NMs to act as transporters of toxic chemicals to sensitive tissues—i.e., **NM Trojan horse effects**.

Once a NM enters a cell, toxicity may occur via one or a combination of these mechanisms. Some toxicity patterns are also emerging; for example, positively charged NMs are generally more toxic than negatively charged NMs (Bexiga 2011; Pagnout 2012; Zheng 2013), although this is not always the case (Lee 2013; Merhi 2012).

Most physico-chemical properties from Table 1 are somehow related to toxicity in several of the mechanisms, however, the quantitative relationships between these properties and the biological uptake and toxicity are not yet clarified. Some examples of the four aforementioned mechanisms are presented below, with an emphasis on the physico-chemical properties, which have been implicated in direct associations with toxicity.

2.1 NM Dissolution

As a representative example of the first toxicity pathway, dissolution of ZnO NMs and subsequent Zn^{2+} release is known to induce cytotoxicity effects, with the mechanisms having recently been elucidated as reactive oxygen species (ROS) generation and activation of an integrated cytotoxic pathway, which involves intracellular calcium flux, mitochondrial depolarization, and plasma membrane leakage (George 2012). A recent seminal work has shown that ZnO cytotoxicity could be reduced by iron doping, which altered the material matrix to diminish Zn^{2+} release (George 2012). This study showed a workflow for identification of an NM descriptor of toxicological relevance, and also provided a strategy to “design out” the toxicity by Fe doping in order to reduce the ZnO NM dissolution potential.

While coating is a typical approach to slow or prevent NM dissolution, recent work has suggested that surface coating or passivation may itself affect the NM core, with more fundamental consequences for stability and toxicity. Thus, surface passivation of 8 nm cobalt ferrite NMs, besides the formation of an iron-rich surface layer, was observed to improve the crystal quality while altering the Fe/Co cation distribution and the NM dissolution rate profile (Soler 2007). Magnetic data revealed that the saturation magnetization increased for surface-passivated NMs compared to the non-passivated ones, though coercivity decreased after passivation. These two phenomena occurred due to changes in the cation distribution among the available tetrahedral and octahedral sites (Soler 2007). Another important outcome from this study is that all of the aged magnetic fluid NM samples deviated from the precursor stoichiometry, thus revealing a discrepancy in the dissolution of cobalt ferrite NMs, which influenced the lattice parameter value (Soler 2007). This highlights the fact that *change of one parameter may inadvertently affect another (or several other) parameter(s)* with direct consequences for quantitative property-activity relationships.

NM composition is the main determinant of whether a NM will dissolve or not, which is linked to the elements' solubility in water. However, as shown in Table 2, particle size and size distribution, which are linked to particle surface area, also influence the rate of dissolution, which typically occurs from the surface. Crystal

Table 2 Contributions of various physico-chemical properties (from Table 1) to the different toxicity mechanisms described here

Measured parameters	Dissolution	NM interactions	NM (surface) specific effects	Trojan Horse
Size/size distribution	✓✓	✓	✓	✓✓
Surface area	✓✓	✓✓	✓✓	✓✓
Purity (particle/dispersant)	~	✓	–	–
Photochemical activity	✓	✓	✓	–
Surface charge/chemistry	✓	✓✓	✓	✓✓
Hydrophobicity	~	✓✓	–	✓✓
Redox activity	✓	–	✓	–
Shape	✓✓	✓	✓✓	✓
Crystal structure	✓✓	✓✓	✓✓	✓✓
Porosity/surface defects	✓	✓	✓✓	✓

Two ticks indicate strong contribution, one tick indicates some contribution, ~ indicates not clear as yet while – indicates likely no significant contribution. Note that these are opinions rather than quantitative values. Measuring the relative contributions quantitatively is challenging and has yet to be achieved

structure and phase also play a role, with faces having coordination numbers $\{1\ 1\ 1\}$ and $\{1\ 1\ 0\}$ dissolving faster than others such as the $\{1\ 0\ 0\}$ face of nanocrystals (Liu 2008; Misra 2012). Shape is also an important parameter, again linked to surface area, but also as narrow areas will dissolve faster than thicker ones. To a lesser degree (although this has not been quantified but is rather an opinion) porosity, again linked to surface area, as well as photochemical and redox activity will drive dissolution, as most cases of nanoparticle (NP) dissolution under environmental conditions are driven by oxidative processes (Ho 2010).

2.2 NM Interactions

2.2.1 Cationic Surface Charge (as Determined by Zeta Potential) Linked to Membrane Damage

The toxic mechanism of a 60-nm cationic (amino-functionalized) polystyrene NM (PS-NH₂) in bacterial cells was explored using a genome-wide collection of bacterial single-gene deletion mutants (Ivask 2012). Over 4000 single nonessential mutants of *Escherichia coli* were screened for the growth phenotype of each strain in the presence and absence of PS-NH₂. The largest number of genes contributing to bacterial sensitivity for PS-NH₂ nanospheres was associated with the formation and functioning of the bacterial cell membrane, followed by defects in lipopolysaccharide and ubiquinone biosynthesis, flagellar formation, and DNA repair. The authors assumed that the proper formation, stability, and functionality of the bacterial cell wall are necessary for the bacteria's ability to compete against cationic NM-induced stress (Ivask 2012). ROS production was also shown to be a crucial pathway of toxicity for PS-NH₂; importantly, it was observed that there is at least one mutant for which there is an additive effect between ROS sensitivity and disruption of membrane integrity. This finding indicates that these two toxicity mechanisms are independent (i.e., disruption of membrane integrity is not necessarily due to ROS production by the NM) (Ivask 2012).

As indicated in Table 2, surface charge thus plays an important role in determining NM interactions with biological and environmental macromolecules, including those incorporated into key biological membranes. Here, hydrophobicity and redox activity will also play a key role in moderating the observed toxicity, with hydrophobicity helping to anchor the NMs close to membranes, and redox activity providing a supply of ROS to amplify the damage from the NMs.

2.2.2 Point of Zero Charge (PZC) and Critical Coagulation Concentration (CCC)

A useful alternative to zeta potential (ζ -potential) measurements [which are very often poorly presented and misinterpreted in the literature, as zeta potential depends

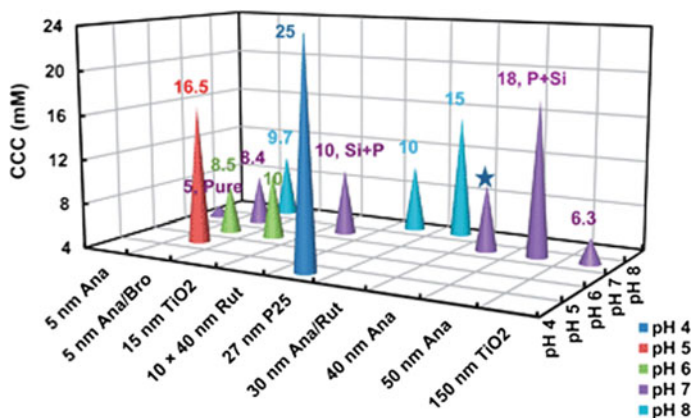


Fig. 3 Comparison of CCC for the TiO₂ NMs with different crystallinity, morphology, and composition at various pH values as reported in the literature (from Liu 2013b and references therein). The CCC values at acidic pH (e.g., pH 5) were generally higher than at neutral pH (e.g., pH 7), which is close to the PZC. The CCC for 10 × 40 nm rutile (10 mM) and 50 nm anatase (18 mM) were higher compared with the 5 nm anatase (5 mM), due to the detected impurities of Si and P and consequently more negative surface charge at pH 7. Further analysis on the material properties of the other TiO₂ is not feasible, due to the insufficient information on the composition of the pristine TiO₂. The *star symbol* stands for the estimated value of CCC, due to the unavailability of the aggregation kinetics data. *Ana* anatase, *Rut* rutile, *Bro* brookite. From Liu (2013b)

on ionic strength, pH and agglomeration/aggregation (Lowry 2016)] is to take into account the context-dependent parameters, such as the point of zero charge (PZC) and the critical coagulation concentration (CCC), since these properties require experimentation under titrating conditions, therefore having less room for dangerous misunderstandings.

An elegant approach to evaluate which physico-chemical properties of TiO₂ NMs and carbon nanotubes determine their environmental stability and transport was recently proposed, although full implementation was prevented by the lack of relevant data under realistic exposure conditions (Liu 2013b). This study showed a high correlation among impurities (including Si and P) resulting from the synthesis route and the PZC, which was in turn correlated with the CCC (Fig. 3) and some implications for transport were discussed. It was suggested that in order to diminish the environmental risks of TiO₂ NMs alternative procedure or chemicals that are Si- and P-free are preferable to use during synthesis (Liu 2013b).

However, in the environment, the adsorption of natural organic matter (NOM) can significantly influence the surface properties and behaviour of TiO₂ NMs (from Liu 2013a and references therein). Thus, further studies are needed regarding the interactions between NOM and TiO₂ of varying properties, as well as to determine whether the property of adsorbed NOM or TiO₂ mainly controls the stability and transport in the natural environment (Liu 2013a).

Other physico-chemical properties strongly linked with NM stability and thus exposure dose and resulting toxicity are photochemical activity, in so much as this can alter the surface composition and thus the PZC and CCC, and likely also crystal structure.

2.2.3 Biological Surface Adsorption Index (BSAI)

A biological surface adsorption index (BSAI) has been proposed as a way to describe the interactions between NMs and biomolecules by estimating the competitive adsorption of a set of small molecule probes onto the NMs (Fig. 4). This is achieved by mimicking the molecular interactions of the NM with the protein residues (Xia 2010). The adsorption of NMs is assumed to depend on Coulomb forces (charged particles), London dispersion (hydrophobic interactions), hydrogen-bond (HB) interactions, dipolarity/polarizability, and lone-pair electrons. Adsorption coefficients of the probe compounds were measured and then were used to construct a set of nanodescriptors representing the contributions and relative strengths of each molecular interaction. The method successfully predicted the

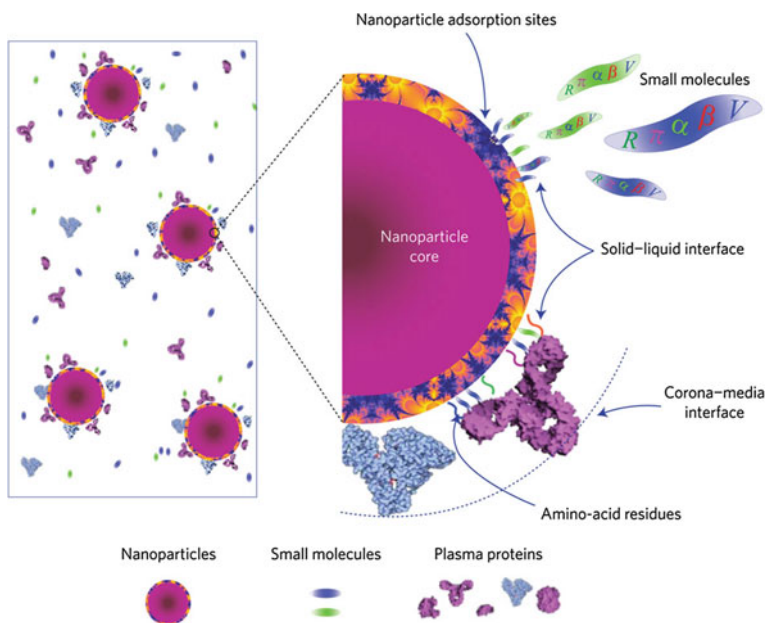


Fig. 4 *Left* in a physiological environment, NMs are exposed to various proteins and small molecules. *Right* the competitive adsorption of small molecules (*upper*) and the residues of proteins (*lower*) on an NM. The *orange ring* on the NM with blue irregular shapes represents the adsorption sites that are not uniformly distributed on the surface. Small molecules with known molecular descriptors $[R, \pi, \alpha, \beta, V]$ can be used as probes to measure the molecular interaction strengths of the NMs with small molecules and biomolecules. From Xia (2010)

adsorption of various small molecules onto carbon nanotubes, and the nanodescriptors were also calculated for 12 other NMs (Xia 2010).

The adsorption coefficients of the probe compounds on a given NM (e.g., multi-walled carbon nanotubes—MWCNTs) were measured using a solid-phase microextraction (SPME)–gas chromatography mass spectrometry (GC–MS) method. The correlation of $\log k$ with the solute descriptors was established after multiple linear regression analysis of the $[\log k, R, \pi, \alpha, \beta, V]$ matrix, where $\log k$ is the adsorption coefficients of a probe compound (that binds to the MWCNT) and $[R, \pi, \alpha, \beta$ and $V]$ are solvation descriptors of the probe compounds—Coulomb forces, London dispersion, hydrogen-bond acidity and basicity, polarizability and lone-pair electrons (Xia 2010). Thus, the BSAI approach provides rational interpretations for the molecular interactions, and also yields five physico-chemical parameters, which characterize the relative strengths of the molecular interactions of the NMs (Xia 2010).

The BSAI nanodescriptors can be related to membrane interaction and biodistribution properties (e.g., absorption rate, distribution coefficient and extent of cellular uptake) of the NM to develop physiologically based pharmacokinetic models, and also for quantitative risk assessment and safety evaluation of NMs (Xia 2010).

Other key physico-chemical parameters affecting biomolecule binding, and thus conferring a biological identity and influencing NM toxicity, include hydrophobicity, surface area and surface curvature (linked to NM size), and crystal structure, as indicated in Table 2. Different NM crystal faces have different energies, resulting in different binding affinities for biomolecules (Lynch et al. 2014a; Sund 2011).

2.2.4 NM-Binding Induced Changes in Protein Conformation May Lead to Receptor Activation

A series of studies using negatively charged poly(acrylic acid)-conjugated gold NMs of various sizes has shown that particle size may affect protein structural changes resulting from binding, which can then induce different modes of interaction between NMs and cells or tissues (Deng 2011, 2013). Activation of the integrin receptor (Mac-1) by 5 nm poly(acrylic acid)-conjugated gold NMs was found to occur due to conformational changes of the bound fibrinogen, leading to increased nuclear factor NF- κ B signalling, which in turn resulted in the release of inflammatory cytokines (Deng 2011). However, larger [20 nm poly(acrylic acid)-conjugated] gold NMs, which also bound fibrinogen, did not induce this effect. This is a *clear demonstration of a NM-protein binding-induced signalling pathway and suggests an alternative mechanism to the more commonly described role of oxidative stress in the inflammatory response to NMs.*

A follow-on study to estimate the effect of binding to gold NMs of different size (5–20 nm) with different surface charge on fibrinogen conformation showed that fibrinogen bound with high affinity to both (positively and negatively charged) NM (Deng 2013). However, binding kinetics and protease digestion suggested that each

NM adopted a different binding orientation, and verified that only the negatively charged NMs induced cytokine release from THP-1 cells (Deng 2011). It was concluded that “*since common proteins can bind to different NMs with quite different biological outcomes, knowledge of the composition of the protein corona is not sufficient to predict biological effects of NMs, and conformational and orientational information is also required*”.

Other NM factors known to influence protein unfolding and formation of so-called cryptic epitopes which can lead to novel toxicities (Lynch 2007), include hydrophobicity, surface curvature [linked to NM size (Klein 2007)] and porosity or surface defects (Clemments 2015).

2.3 NM (Surface) Specific Effects

2.3.1 Surface Defects, Including Those Induced by Surface Oxidation

A study of the effect of nanosize Ag spheres, plates, and wires on a fish gill epithelial cell line (RT-W1) and on zebrafish embryos showed significantly increased toxicity from the Ag nanoplates compared to the other particle shapes (George 2012). Features, such as Ag ion shedding and bioavailability failed to thoroughly explain the enhanced toxicity of the nanoplates. High-resolution transmission electron microscopy showed a high level of crystal defects (stacking faults and point defects) on the nanoplate surfaces. A noticeable reduction of toxicity in RT-W1 cells and zebrafish embryos was observed upon surface coating with cysteine to passivate the surface defects (George 2012).

Graphene oxide (GO) is considered as being biocompatible, but until recently there has been a limited amount of data to verify this. Currently, four general top-down routes for GO production use different acids to oxidise the surface and offer hydrophilicity. The cytotoxicity of GOs (prepared by the four oxidative treatments above) was measured by means of the mitochondrial activity in adherent lung epithelial cells (A549), using commercially available viability assays (MTT and WST-8) (Chng 2013). All four GO NMs yielded strong dose-dependent cytotoxic responses after 24 h exposure, and a relation between the oxygen content/functional groups of GOs with their toxicological response against the A549 cells was observed: The various oxidative approaches produced GOs with different properties due to varying C/O ratios and proportions of the types of oxygen-containing groups (e.g., carbonyl group) (Chng 2013). *This is the first study, which demonstrates how the oxidative routes employed to prepare GOs (also other carbon-based NMs) may profoundly affect their toxicity.*

Another article demonstrating that the selected method to vary the NM physico-chemical features can itself induce other unwanted effects is that of Bussy et al. (2012), where the impact of carbon nanotube length on toxicity was investigated (Bussy 2012). A broad study was designed to compare the effects of two samples of MWCNT (synthesized following a similar production process, i.e.,

aerosol-assisted CVD) on murine macrophages; a soft ultrasonic treatment in water was used to change the length of one of the MWCNTs (Bussy 2012). It was shown that altering the length of MWCNT leads to associated structural (i.e., defects) and chemical (i.e., oxidation) changes that affect both surface and residual catalyst iron NM content of the CNT. The structural defects and oxidation (induced by the length reduction process) were shown to be at least as responsible as the length reduction itself for the increased pro-inflammatory and pro-oxidative response observed with short (oxidized) MWCNT compared to long (pristine) MWCNT (Bussy 2012). This further demonstrates the problem of the inter-dependence of various physico-chemical properties and the *significant challenges inherent in the development of systematically varied libraries of NMs as the basis for mechanistic studies*. However, designing libraries with this knowledge of predicted physico-chemical interdependencies in place will allow this information to be included into the physico-chemical testing strategy and into the structure-property relationships under development.

2.3.2 Bandgap as a Proxy for Oxidative Stress

Oxide semiconductors can serve as channels for electron transfer between aqueous reactants. The occurrence of these transfers depends on similarities in the energetic states of the NMs and ambient redox-active aqueous substances (Zhang 2012b). Burello and Worth proposed a theoretical scheme where the relationship between the cellular redox potential to metal oxide (MOx) band gap clarifies the observed oxidative stress and toxicity generation by some of these materials (Burello 2011). According to this band gap hypothesis, it is possible to predict the oxidative stress potential of MOx NMs by comparing the E_v (valence band) and E_c (conduction band) levels to the cellular redox potential (Burello 2011).

Using a panel of 24 MOx NMs (George et al. 2012) showed that with the use of conduction band energy levels (band gap), it is possible to describe their toxicological potential at cellular and whole organism levels. Among the NMs, the overlap of conduction band energy (E_c) levels with the cellular redox potential (-4.12 to -4.84 eV) was strongly associated with the ability of Co_3O_4 , Cr_2O_3 , Ni_2O_3 , Mn_2O_3 and CoO NMs to induce oxygen radicals, oxidative stress, and inflammation (George 2012). Although CuO and ZnO produced oxidative stress and acute pulmonary inflammation, which is not predicted by E_c levels, the adverse biological effects of these NMs are explained by their solubility, as demonstrated by ICP-MS experiments (George 2012). These results indicated that the toxicity of a large series of MOx NMs can be predicted in the lung based on semiconductor properties and an integrated in vitro/in vivo hazard ranking model based on oxidative stress. It is not yet clear whether there are additional factors that contribute to band-gap in addition to crystal structure and composition, although for silicon wires bond angles and bond strain cause a transition from direct to indirect band gap behaviour and the properties can be tailored through surface chemistry (Brus 1994). This could be an area for future nanosafety research, focussing on, for

example, the role of porosity, defects, doping or anion exchange on energy transfer within nanocrystals.

2.4 *NM Trojan Horse Effects—Increased Local Concentrations of (Dissolved) Species*

In medicine, the Trojan horse effect is defined as “Any disastrous result of an anticipated gain; or, the masking of a dangerous agent within an innocent garb”.¹ There are numerous cases where NMs act as Trojan horses, both when the NM alone releases high concentrations of ionic species in sites that they would not normally reach such high concentrations (Park 2010; Studer 2010), and when the NM carries another compound along, thus accessing an otherwise inaccessible location (Baun 2008). Both types of the Trojan horse effect are based on the size properties and biomolecule adsorption of NMs, which enable them to participate in the active receptor-mediated transport processes of cells and organisms that are often less accessible to ionic or molecular species, which are typically internalised based on passive diffusion (Salvati 2011).

Silver NMs (distributed in foetal bovine serum, average size: 68.9 nm, concentrations: 0.2, 0.4, 0.8, and 1.6 ppm, exposure time: 24, 48, 72, and 96 h) appeared cytotoxic to cultured RAW264.7 cells by increasing sub G1 fraction, which denotes cellular apoptosis (Park 2010). Silver NMs were found in the cytosol of activated cells, whereas were absent in the dead cells, thus suggesting their dissolution and release of ions and cytotoxicity by a Trojan-horse type mechanism (Park 2010). Studer et al. (2010) studied the toxicities of copper oxide and carbon coated copper metal NMs at constant copper exposure dose, and observed noticeably different responses from the two NM forms: while copper oxide was highly cytotoxic, carbon-coated copper NMs were much less cytotoxic and more tolerated, which corresponded with the two material's intra- and extracellular solubility in model buffers (Studer 2010). Thus, the differences in toxicity correlated with different copper release in line with a Trojan horse-type scenario.

C₆₀-NMs (Buckminster fullerenes) are known to function as carriers of contaminants in aqueous systems. This has been verified in a series of toxicity tests with algae (*Pseudokirchneriella subcapitata*) and crustaceans (*Daphnia magna*) with four common environmental contaminants (atrazine, methyl parathion, pentachlorophenol (PCP), and phenanthrene) as model compounds with different physico-chemical properties and toxic modes of action (Baun 2008). In algal tests, C₆₀-aggregates increased the toxicity of phenanthrene by 60% and decreased the toxicity of PCP about twofold. Addition of C₆₀-aggregates increased the toxicity of phenanthrene ten fold when results were expressed as water phase concentrations. Metals, such as Cd bound to NM (e.g., titania) surface, show increased bioavailability in fish (carp) (Shaw

¹<http://medical-dictionary.thefreedictionary.com/Trojan+Horse+Effect>.

2008). Since many NMs are negatively charged, they act as suitable carriers for cationic metals. The data mentioned above underline that *both inherent toxicity of manufactured NMs, and interactions with other compounds and characterisation of NMs in aqueous suspension are crucial data for risk assessment of NMs.*

Table 2 also highlighted the range of physiochemical parameters that can play a role in determining pollutant binding to NMs, as a prerequisite for Trojan horse effects resulting from the pollutant gaining a new access route carried on the NM. This also shows that the same NM physico-chemical parameter may contribute to multiple modes of action, which may occur simultaneously.

3 Systematic NM Libraries Reported in the Literature

To associate particular physico-chemical properties of an NM with its toxicity, it is imperative to establish combinatorial libraries, constructed in a way that systematic variation of important physico-chemical properties (probably related to toxicity) can occur (Xia 2012). Property variations may include NM size, shape, surface area, band gap, porosity, crystallinity, charge, solubility, and surface functionalization, as shown in Fig. 5.

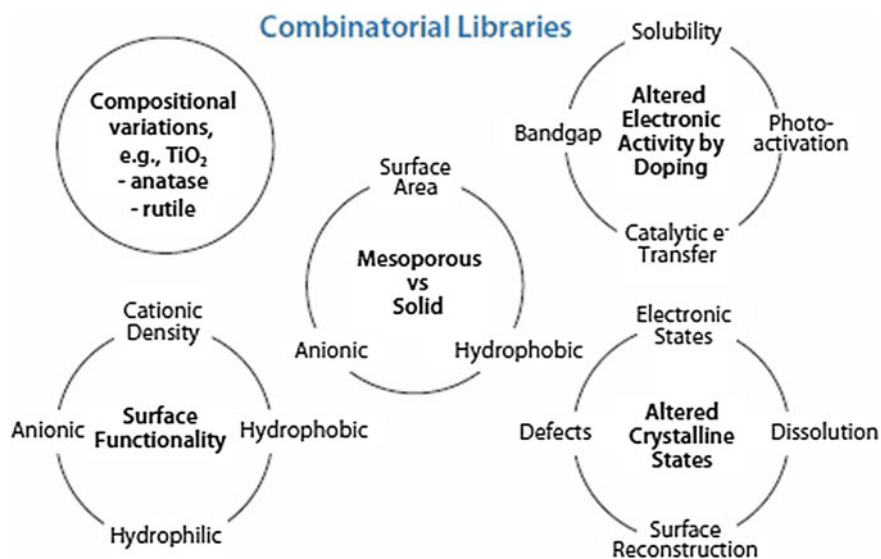


Fig. 5 Examples of combinatorial NM libraries. Combinatorial libraries are constructed by synthesizing one of the compositional NMs to vary physico-chemical properties, which may be involved in toxicity. Property variations apply to NM size, shape, charge, porosity, hydrophilicity, hydrophobicity, crystallinity, band gap, photoactivation, solubility, and surface area. A single property change may also affect other properties, thus rigorous re-characterization is required. From Xia (2012)

Here, we offer an overview of the types of NM libraries that have been developed in academic groups to date, although these are not available commercially, and they have typically varied only one or two materials, thus preventing safe generalisations. Important attempts to develop QNARs based on these libraries are also presented. So far, two tools to address the need for constructing validated QNARs have been developed and are disseminated as ready-to-use applications for anyone interested in NMs risk assessment. These tools aim to investigate the NMs' diverse effects by estimating the impact of different surface modifiers on cellular uptake and by exploring the dependency between various physico-chemical descriptors of iron oxide and toxicity. These two approaches may facilitate ongoing research on the identification of the crucial descriptors for the virtual screening of NMs diverse effects.

One of the few organized datasets of NMs that has been presented in the literature contains the cellular uptake of 109 NMs in pancreatic cancer cells (PaCa2). Each NM within this dataset involves the same metal core (iron oxide/NH₂ cores) but different surface modifiers, which are small organic molecules conjugated to the NM surface (Weissleder 2005). For the development of our first tool, we have constructed and validated a QNSAR model for the prediction of the cellular uptake in pancreatic cancer cells based on this dataset. The *in silico* workflow is available online through the Enalos InSilicoNano platform (http://enalos.insilicotox.com/QNAR_PaCa2/), which is a web service based solely on open source and freely available software that was developed to make our model available to any one aiming at acquiring knowledge on potential biological effects in the decision making framework (Fig. 6). To test the usefulness of the web service, the entire PubChem database was exploited to select compounds similar to a known active structure. Next, the Enalos InSilicoNano platform was used to identify novel potent NMs from a prioritized list of compounds (Melagraki 2014).

Name	Description
QNAR_PaCa2	QNAR model correlating chemical descriptors and MNP cellular uptakes (similar nanoparticle core with different surface modifiers)

Fig. 6 Screenshot of the Enalos Platform input page for the prediction of NMs uptake in PaCa2 cells

A second online tool for the computational screening of iron oxide NMs was also developed to complement our previously reported efforts to extract valuable information from available datasets and to develop user friendly applications for the risk assessment of NMs. For this purpose, a predictive classification model was developed for the toxicological evaluation of 44 iron oxide NMs with different core, coating and surface modifications based on several different properties, such as size, relaxivities, zeta potential and type of coating (Shaw 2008).

The model was fully validated through several validation tests and was released online via the Enalos InSilicoNano Platform (http://enalos.insilicotox.com/QNAR_IronOxide_Toxicity/). This web service allows a user to insert specific properties (Fig. 7) and subsequently to obtain a toxicity prediction (and an indication of the reliability of the prediction) based on the domain of applicability (Melagraki 2015).

 **Enalos QNAR Iron Oxide Toxicity Platform**

MNP Number	Size (nm)	ZP (mV)	R1 (mM-1S-1)	R2 (mM-1S-1)	Coating
1	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>
2	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>
3	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>
4	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>
5	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>
6	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>
7	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>
8	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>
9	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>
10	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>
11	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>
12	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>
13	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>
14	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>
15	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>
16	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>
17	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>
18	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>
19	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>
20	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	Other <input type="button" value="v"/>

Import a CSV file for High Throughput Virtual Screening (.csv)

No file selected.

Fig. 7 Screen shot of Enalos QNAR iron oxide toxicity platform input page

These freely available web services are the first efforts to establish an online tool that the wider scientific community can easily apply in the computer-aided NM design and may be a means to reliably predict the activity of novel nano-structures. The user friendly environment enables different datasets to be directly imported based on the particular requirements of the user. Both web services offered via the Enalos InSilicoNano platform aim to provide significant aid within a virtual screening framework, for the design of novel NMs or their prioritization based on predicted diverse effects.

3.1 Gold Nanoparticle Libraries

Gold nanoparticles were selected by a group at the university of Oregon to represent an ideal platform for the systematic evaluation of the effect of various physico-chemical properties (including core size, charge, and surface chemistry) on biological responses to NM exposure (Harper 2011). A matrix of nine structurally diverse, precision-engineered Au NMs of high purity and known composition was constructed. This included three core sizes and four unique surface coatings that contain positively and negatively charged head groups, as shown in Fig. 8.

Testing the different combinations of core sizes and ligand shells enabled evaluation of the importance of small changes in core size and ligand composition independently. Zebrafish embryo mortality, malformations, uptake, and elimination of Au NMs depended on the above parameters, thus highlighting the need for very careful experimental design and NM characterization (Harper 2011).

Besides the great control over particle size and the simple functionalization through surface modification (Fig. 9), gold NMs are excellent systems for further development of NMs libraries, due to the recent developments in controlling the shape of the nanocrystals (Sau 2009). The size, shape, and structural control of the nanocrystals is achieved through handling of the kinetic and thermodynamic parameters of the systems with the aid of additives, light and thermal energy, as well as their combinations (Sau 2009 and references therein). The formation of diverse shapes most likely arises from the relationship between the faceting tendency of the stabilizing agent and the growth kinetics (rate of supply of Au⁰ to the crystal planes) (Ahmadi 1996). Sau et al. presented a seed-mediated growth method to regulate the morphology and dimensions of gold nanocrystals by manipulating the experimental parameters in aqueous medium at room temperature (Sau 2009). This chemical procedure generated several architectures with rod-, rectangle-, triangle-, hexagon-, cube-, and star like structures and branched (i.e., bi-, tri-, tetra-, and multipod) Au nanocrystals of varied dimensions in high yield, in the presence of a single surfactant (cetyltrimethylammonium bromide, Fig. 8) (Sau 2004). Although this surfactant is not typically used in toxicology studies, this work makes evident that shape maybe changed in a manner such that all shapes result from identical starting structures. This observation is important because it suggests that only the shape changes from one material to another, while other parameters (e.g.,

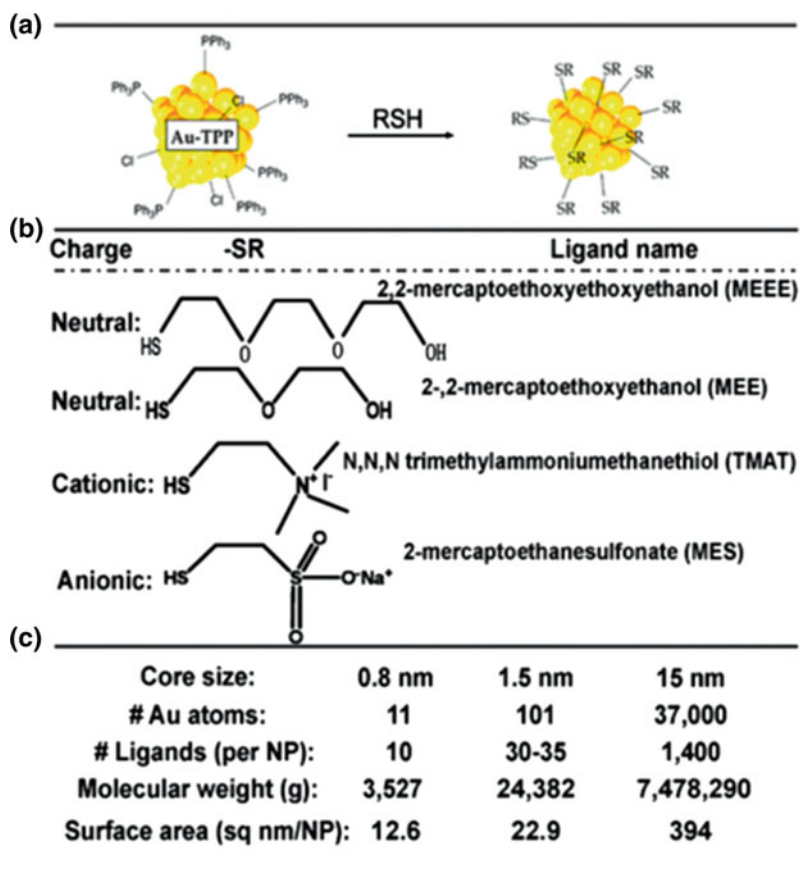


Fig. 8 Gold NM synthesis, structure, and properties. **a** Synthesis reaction of functionalized gold NMs from gold triphenylphosphine (AuTTP) NM building blocks. **b** Name, charge, and structure are provided for each ligand tested. **c** Properties with implications for dose metrics for the different NM core sizes are also given. From Harper (2011)

defects, faceting face etc.) are also obviously affected. Moreover, this approach would allow the *experimental evaluation of differences in biomolecule binding, diffusion and interaction with receptors etc. based on NM shape in a systematic manner. This may constitute the basis for further improvement and validation of models and QSARs.* For example, molecular dynamics (MD) simulations have shown that the efficacy of passive endocytosis is higher for spherocylindrical particles compared to spheres and that endocytosis is abolished for sharp-edge particles (Vácha 2011).

Walkey et al. have recently presented the synthesis and experimental evaluation of a chemically diverse set of 105 gold NMs with various surface modifiers (Walkey 2014). The authors considered three different core sizes, namely 15, 30

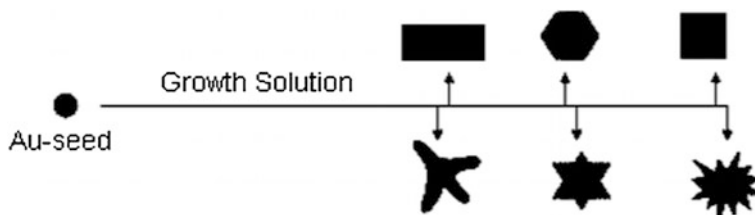


Fig. 9 Schematic illustration of the variation of gold NM morphology and dimensions (shape and structure) by manipulation of the experimental parameters in aqueous solution at room temperature. From Sau (2004)

and 60 nm, and employed 67 organic surface modifiers, which included small molecules, polymers, peptides, surfactants and lipids that can be identified as “neutral”, “anionic” and “cationic” according to their structural and charge properties at physiological pH. Different parameters for each NM were measured before and after exposure in blood serum stimulating the biomolecular environment during *in vitro* cell culture experiments. In total, 21 physico-chemical parameters before and after exposure were calculated and 785 distinct serum proteins were recognized on the NMs. Furthermore, all NMs were evaluated regarding their association with A549 human lung epithelial carcinoma cells (Albanese 2014).

3.2 Metal Oxide Nanoparticle Libraries

The Centre for Environmental Implications of Nanoscience and Nanotechnologies at UCLA have developed a compositional library of 24 different metal oxide NMs (whose compositions, size distribution and zeta potential are shown in Table 3), which was constructed to explore the correlation of NM electronic properties and their toxicity results (Lin et al. 2013). Most of the NMs were directly purchased, and others (CuO, Co₃O₄, Fe₃O₄, Sb₂O₃, TiO₂, WO₃, and ZnO) were synthesized in-house by flame spray pyrolysis (Teoh 2010). The results of Table 3 suggested that the hydrodynamic size of these NMs in Holtfreter’s medium (pH 7.0) ranges from 200 to 500 nm, with only a few materials (Al₂O₃, Fe₃O₄, Ni₂O₃, SnO₂, Y₂O₃, Yb₂O₃) reaching 500 nm. The ζ -potentials of all NMs were negative (–20 to –30 mV) due to alginate coating.

Using the above NM library, Lin et al. (2013) established a predictive toxicological paradigm (that of metal oxide dissolution and ligation of the zebrafish hatching enzyme 1 (ZHE1) enzyme centre by specific metal ions), which can be used for the safety evaluation of dissolved metal oxide NMs in aqueous media. The excellent correlation between ZHE1 inactivation and hatching interference in intact embryos rationalized the molecular mechanism of hatching interference exerted by CuO, ZnO, Cr₂O₃ and NiO NMs (Lin et al. 2013). The authors concluded that shedding of metal ions by dissolvable metal oxide NMs interferes with recombinant

Table 3 Physico-chemical features of 24 metal oxides in the UCLA library

Number	MOx	Primary size (nm) ^a	Hydrodynamic size (nm) ^b	Zeta-potential (mV) ^c
1	Al ₂ O ₃	14.7 ± 5.2	524.8 ± 32.8	-24.0 ± 0.5
2	CeO ₂	12.8 ± 3.4	321.3 ± 8.6	-28.9 ± 3.3
3	CoO	18.3 ± 6.8	378.3 ± 16.4	-25.5 ± 1.3
4	Co ₃ O ₄	10.0 ± 2.4	247.6 ± 16.9	-29.0 ± 2.2
5	Cr ₂ O ₃	71.8 ± 16.2	478.5 ± 7.2	-26.2 ± 3.1
6	CuO	193.0 ± 90.0	289.5 ± 31.0	-26.9 ± 0.8
7	Fe ₂ O ₃	12.3 ± 2.9	385.2 ± 6.3	-24.1 ± 2.0
8	Fe ₃ O ₄	12.0 ± 3.2	831.7 ± 41.8	-27.0 ± 2.3
9	Gd ₂ O ₃	43.8 ± 15.8	726.7 ± 54.8	-34.7 ± 0.7
10	HfO ₂	28.4 ± 7.3	349.9 ± 5.2	-24.3 ± 2.1
11	In ₂ O ₃	59.6 ± 19.0	303.2 ± 5.2	-35.5 ± 2.4
12	La ₂ O ₃	24.6 ± 5.3	471.2 ± 20.9	-27.8 ± 0.6
13	Mn ₂ O ₃	51.5 ± 7.3	525.9 ± 7.8	-30.9 ± 0.4
14	NiO	13.1 ± 5.9	277.5 ± 23.0	-23.1 ± 2.0
15	Ni ₂ O ₃	140.6 ± 52.5	665.8 ± 46.4	-24.4 ± 2.2
16	Sb ₂ O ₃	11.8 ± 3.3	459.9 ± 22.7	-25.8 ± 0.9
17	SiO ₂	13.5 ± 4.2	374.9 ± 29.0	-16.8 ± 2.0
18	SnO ₂	62.4 ± 13.2	635.0 ± 52.0	-26.4 ± 0.3
19	TiO ₂	12.6 ± 4.3	497.0 ± 17.1	-31.5 ± 1.4
20	WO ₂	16.6 ± 4.3	511.9 ± 19.4	-23.3 ± 1.1
21	Y ₂ O ₃	32.7 ± 8.1	594.5 ± 33.0	-27.6 ± 0.4
22	Yb ₂ O ₃	61.7 ± 11.3	682.6 ± 56.2	-29.7 ± 0.5
23	ZnO	22.6 ± 5.1	379.0 ± 11.0	-27.0 ± 1.1
24	ZrO ₂	40.1 ± 12.6	384.4 ± 25.0	-19.7 ± 3.6

From Lin et al. (2013)

^aPrimary size of particles in their dry state was obtained by transmission electron microscopy (JEOL, 1200 EX)

^bHydrodynamic size was determined by high throughput dynamic light scattering (HT-DLS, Dynapro Plate Reader, Wyatt Tech)

^cParticle ζ-potential was measured using ZetaPALS (Brookhaven Instruments, Holtsville, NY). Introduction of the NPs in Holtfreter's medium (pH 7.0) did not significantly change the medium pH in spite of the dissolution of metal oxide NPs

ZHE1 activity. Consequently, it was anticipated that CuO, ZnO, Cr₂O₃ and NiO NMs should interfere in embryo hatching (Lin et al. 2013).

A classification-based cytotoxicity nanostructure-activity relationship (nano-SAR) with excellent classification accuracy was developed based on four descriptors: atomization energy of the metal oxide, period of the NM metal, NM primary size, and NM volume fraction (in solution). Based on a set of 9 metal oxide NMs to which transformed bronchial epithelial cells (BEAS-2B) were exposed over a range of concentrations and exposure times up to 24 h, the best-performing model had a 100% classification accuracy in both internal and external validation (Liu 2011).

Additional applications of this NM library include the evaluation of a possible connection between conduction band energy levels and toxicity, specifically oxidative stress and acute pulmonary inflammation (Zhang 2012b). The authors employed the same set of NMs from Table 3 to prove that the overlap of conduction band energy (E_c) levels with the cellular redox potential (-4.12 to -4.84 eV) was highly correlated to the ability of Co_3O_4 , Mn_2O_3 , Cr_2O_3 , Ni_2O_3 and CoO NMs to induce oxygen radicals, oxidative stress, and inflammation (Zhang 2012b).

Recently, the UCLA Centre have also used the flame spray pyrolysis (FSP) technique to construct NM libraries where metal oxides are doped with Fe to diminish their solubility or toxicity (e.g., pure or Fe-doped ZnO or TiO_2 NMs) (Pokhrel 2012). It was demonstrated that FSP is a flexible method for the effective design of a homologous library (i.e., a library based on a parent oxide, which is doped with various amounts of dopant) and may also serve as a tool for exploring the properties of the resulting compounds (Pokhrel 2012).

Another NM library of 17 metal oxides (size range: 15–90 nm) was established by Puzyn et al. (2011). Experimental and computational results related to the toxicity of the NMs in terms of predicted EC_{50} values (calculated using a single descriptor, ΔH_{Me^+} describing ionization enthalpy of the detached metal atoms) are shown in Table 4 (Puzyn et al. 2011). Since particle size does not significantly affect toxicity in the above size range, the chosen descriptors essentially represent reactivity-related electronic properties. The NMs, which were used in the training set to develop the QSAR equation are denoted by T (Table 4), and the NMs in the validation sets by V_1 and V_2 . The model reliably predicted the toxicity of all compounds under study (Puzyn et al. 2011).

3.3 Silica Nanoparticle Library

A silica NM library consisting of 12 variants of amorphous Stöber silica, mesoporous silica, amorphous fumed silica, silicalite and α -quartz was constructed to investigate the crystallinity and surface effects of silica NMs (Zhang 2012a). Particular physico-chemical properties, such as size, size distribution and zeta potential in water and two different media (containing protein or serum) of the different NMs are presented in Table 5. Initial studies indicated that fumed silica is the most toxic among the silicas in the library (Zhang 2012a). The toxicity can be reduced by heating and re-imposed by hydration. The changes are most likely related to the number of surface silanol groups and strained three-member Si rings (Zhang 2012a).

Table 4 Experimental and predicted features of metal oxides used in the Puzyn study

Metal oxide	Descriptor	Leverage value, h	Observed log $1/EC_{50}$ (mol l^{-1})	Predicted log $1/EC_{50}$ (mol l^{-1})	Residuals	Set
	ΔH_{Me+} (kcal mol^{-1})					
ZnO	662.44	0.33	3.45	3.30	0.15	T
CuO	706.25	0.29	3.20	3.24	-0.04	T
V ₂ O ₃	1,097.73	0.11	3.14	2.74	0.40	V_1
Y ₂ O ₃	837.15	0.21	2.87	3.08	-0.21	T
Bi ₂ O ₃	1,137.40	0.10	2.82	2.69	0.13	T
In ₂ O ₃	1,271.13	0.10	2.81	2.52	0.29	T
Sb ₂ O ₃	1,233.06	0.10	2.64	2.57	0.07	V_1
Al ₂ O ₃	1,187.83	0.10	2.49	2.63	-0.14	T
Fe ₂ O ₃	1,408.29	0.13	2.29	2.35	-0.06	T
SiO ₂	1,686.38	0.26	2.20	1.99	0.21	T
ZrO ₂	1,357.66	0.11	2.15	2.41	-0.26	V_1
SnO ₂	1,717.32	0.28	2.01	1.95	0.06	T
TiO ₂	1,575.73	0.19	1.74	2.13	-0.39	T
CoO	601.80	0.38	3.51	3.38	0.13	V_2
NiO	596.70	0.39	3.45	3.38	0.07	V_2
Cr ₂ O ₃	1,268.70	0.10	2.51	2.52	-0.01	V_2
La ₂ O ₃	1,017.22	0.13	2.87	2.85	0.02	V_2

From Puzyn et al. (2011)

The leverage value h (acceptable if not higher than 0.6) indicates deviations of the structure of the compound from those used for the QSAR development

3.4 Single-Wall Carbon Nanotube Library

A three-particle (HiPco-D, SG65-D, and P2-D), single-walled carbon nanotube (SWCNT) library was created to explore the impact of hydrophobicity, metal impurity, and dispersion state on NM toxicity (Chowdhury et al. 2012). In Table 6, the physico-chemical features and the metal contaminants are presented. Using these three SWCNTs, Chowdhury et al. (2012) initiated a study to correlate their transport with the various synthetic methods and residual catalyst contents in order to elucidate their effect on the nanotubes (Chowdhury et al. 2012). After purification, the residual metal catalyst between the SWCNTs follows the trend: HiPco-D > SG65-D > P2-D. The electrophoretic mobility (EPM) and hydrodynamic diameter of SWCNTs remained unaffected by SWCNT type, pH, and presence of natural organic matter (NOM); nevertheless, the ionic strength (IS) and ion valence (K^+ , Ca^{2+}) altered the hydrodynamic diameter and EPM properties of SWCNTs. Overall, it was concluded that the different synthetic approaches resulted in unique breakthrough trends, which were related to metal content (Chowdhury et al. 2012).

Table 5 Physico-chemical characteristics of silica NMs in the UCLA library

Silica NM	Size (nm)			Zeta potential (mV)		
	H ₂ O	BEGM (2 mg/mL BSA)	DMEM (10% FCS)	H ₂ O	BEGM (2 mg/mL BSA)	DMEM (10% FCS)
As prepared Stöber silica	118.4 ± 1.9	257.6 ± 22.6	196.7 ± 11.7	-39.2 ± 1.4	-5.5 ± 3.5	-25.2 ± 14.1
Stöbercalcinated at 600 °C	156.6 ± 3.2	282.6 ± 3.2	227.8 ± 0.9	-26.8 ± 0.5	-8.5 ± 4.1	-10.9 ± 2.4
Stöbercalcinated at 800 °C	211.0 ± 6.2	311.9 ± 7.1	233.2 ± 3.6	-34.9 ± 1.6	-10.7 ± 2.8	-5.7 ± 2.8
Rehydrated Stöber silica	143.3 ± 3.8	249.9 ± 2.2	203.6 ± 4.7	-36.8 ± 2	-9.6 ± 4.4	-11.0 ± 6.9
Aggregated Stöber silica	148.1 ± 2.1	532.8 ± 27.8	231.2 ± 2.3	-8.81 ± 0.1	-9.04 ± 1.7	-11.24 ± 1.5
As received fumed silica	131.5 ± 0.6	627.6 ± 69.9	256.3 ± 19	-22.4 ± 7.1	-4.5 ± 3.1	-10.1 ± 8.5
Fumed silica calcinated at 600 °C	286.2 ± 22.5	667.3 ± 23.6	326.3 ± 3.9	-33.2 ± 3.7	-11.5 ± 4.9	-4.9 ± 2.2
Fumed silica calcinated at 800 °C	310.1 ± 11.2	676.7 ± 32.8	339.0 ± 9.9	-19.0 ± 1.6	-7.1 ± 1.2	-4.1 ± 1.7
Rehydrated fumed silica	161.6 ± 29	623.3 ± 18.6	264.4 ± 7.3	-40.5 ± 2.2	-10.7 ± 1.5	-6.5 ± 2.1
Aggregated fumed silica	263.3 ± 7.8	409.3 ± 43.2	248.6 ± 12.8	-58.0 ± 1.3	-0.9 ± 7.9	-2.1 ± 10.9
Mesoporous silica	235.0 ± 7.2	364.8 ± 128.4	362.9 ± 51.1	-43.9 ± 3.6	-18.4 ± 8.7	-4.8 ± 13.1
Min-U-Sil quartz	94.3 ± 5.5	246.1 ± 35.9	126.8 ± 8.0	-15.5 ± 0.3	-19.3 ± 1.6	-7.74 ± 0.3

From Zhang et al. (2012a, b)

Table 6 Physical Characterization of different single-walled carbon nanotubes

	Diameter range (nm) ^a	Length (nm) ^b	Metal mass (%) ^c	Species ^c	Species mass (%) ^c	Metal species (%) ^c
HiPco-D	0.696–1.129	383 ± 281	6.52	Fe	6.52	100
SG65-D	0.682–0.981	449 ± 316	1.80	Co	0.22	12.26
				Mo	1.58	87.74
P2-D	1.158–1.699	404 ± 221	0.21	Y	0.06	30.24
				Ni	0.15	69.76

From Chowdhury et al. (2012)

^aThe diameters of SWCNTs are determined by vis-NIR and TEM

^bLengths of SWCNTs were determined from atomic force microscopy

^cResidual metal species and mass determined via inductively coupled plasma atomic emission spectroscopy

3.5 Combinatorial Synthesis of (Biodegradable) Polymers

Intensive research related to nanomedicine and nanosafety assessment is also evolving. For instance, combinatorial techniques are widely used in the pharmaceutical industry, and are increasingly being applied to the development of polymer coatings. A combinatorial synthesis of biodegradable polyanhydride film and NM libraries as well as the high-throughput detection of protein release from these libraries has recently been reported (Petersen 2012). The method enables the rapid construction of micro-scale polymer libraries, reducing the batch size while creating multivariant polymer systems. Moreover, the combinatorial polymer library can be fabricated into blank or protein-loaded geometries upon dissolution of the polymer library in a solvent and precipitation into a non-solvent (for NMs) or by vacuum drying (for films). The libraries have been screened for protein release kinetics, stability and antigenicity; in vitro cellular toxicity, cytokine production, surface marker expression, adhesion, proliferation and differentiation; and in vivo biodistribution and mucoadhesion (Petersen 2012 and references therein). Such approaches are very promising regarding the development of systematically varied NM libraries and their evaluation.

4 Gap Between Measured Physico-chemical Parameters and Calculated QSAR Descriptors

Given the dynamic nature of NMs, and their context-dependence, there is a disconnect between the physico-chemical parameters characterised routinely and those utilised in QSAR models (Valsami-Jones 2015). Indeed there is still considerable variability in the various lists of minimal characterisation parameters that have been discussed over the last 10 years (see Stefaniak 2013 for a review of these lists) and their degree of overlap. A recent editorial in ACS Nano described this succinctly:

“Structure-activity analyses of well-characterized material libraries (used for exploring a series of nano/bio interfaces) have elucidated nanoscale-specific properties that go beyond the traditional lists of intrinsic and extrinsic property characterization” (Nel 2015). Examples of novel or non-traditional characteristics or descriptors linked to novel toxicities include band gap and hydration energies which have been linked to generation of oxidative stress in bacteria and mammalian cells by metal oxide semiconductor materials (Zhang 2012b); surface strain resulting from highly reactive silanols leading to membrane damage by pyrolytic silica (Zhang 2012a), or unfolding of proteins at the NM surface leading to activation of inflammatory receptors (Deng 2011). Examples related to the occurrence of dynamic changes or interactions (which have been termed extrinsic effects) include the complexation of structural cellular phosphate residues on the surface of rare earth oxides and up-conversion NP leading to lysosome damage (Li 2014); and the degradation of surface coatings attached to NMs in the acidic lysosomal environment, leading to lysosomal injury (Wang 2015). As pointed out by Nel et al., none of these structure-activity relationships could have been predicted using the traditional list of intrinsic and extrinsic property evaluations (Nel 2015). To further illustrate this point, we have reviewed the QSAR literature and extracted the descriptors that were used as the basis of predictive models compared to those physico-chemical parameters of the NMs that were characterised, with very clear divergence and discrepancies apparent (See Table 7). Thus, we either need to develop models utilising the easily measured physico-chemical parameters, find ways to extract more useful data from the traditional methods, or develop approaches to routinely measure those parameters more directly linked to structure-activity relationships.

As evident from this snapshot of QSAR models for NMs, most models are based on calculated parameters that take no or limited account of the NMs characteristics in the exposure medium or are derived from crystallographic data (for example) (Puzyn et al. 2011) and are certainly not reliant on detailed physico-chemical characterisation of the dynamics of NM behaviour in the biological medium and organism. More recent efforts include extraction of detailed parameters from TEM images (e.g. Gajewicz 2015) or protein adsorption to the NMs (e.g. Liu 2015a) which are experimentally determined parameters, although corona characterisation is not yet a routine characterisation or required as part of regulatory dossiers for NMs, in large part due to the cost involved and the early stages in terms of verification of a predictive relationship for NM uptake/toxicity. Figure 10 shows an example utilising a calculated version of acid-base properties, which increased with the number of oxygens in the oxide NM and proportion of surface molecules to molecules in volume as descriptors to model NM toxicity to bacteria (Sizochenko 2014). As available datasets increase in number and size, models comparing increasingly diverse NMs and exposure conditions are becoming available, suggesting that significant progress is being made in this arena, although there is still a need for concerted effort in this research field.

It is clear that there is a need for development of a strategy for identification of NMs' critical properties linked to biological impacts, and in particular to tease out

Table 7 Summary of selected QSAR studies from the literature highlighting the physico-chemical parameters measured and the calculated descriptors

NMs used	End-point and biological system	Physico-chemical properties characterised	Calculated descriptors	References
17 nano-metal oxides	Toxicity to <i>E. coli</i> bacteria	No NM characterisation Crystallographic data used to calculate 12 structural characteristics EC ₅₀ determined experimentally and calculated from model	<p>$\Delta H_{Me^{n+}}$, which represents enthalpy of formation of a gaseous cation having the same oxidation state as that in the metal oxide structure</p> <p>Structural parameters:</p> <ul style="list-style-type: none"> • Standard heat of formation of the oxide cluster • Total energy of the oxide cluster • Electronic energy of oxide cluster • Core-core repulsion energy of oxide cluster • Area of oxide cluster calculated based on COSMO • Volume of oxide cluster calculated based on COSMO • Energy of HOMO • Energy of LUMO • Energy difference between HOMO and LUMO • Enthalpy of detachment of metal cations(Meⁿ⁺) from cluster surface • Enthalpy of formation of a gaseous cation • Lattice energy of oxide 	Puzyn et al. (2011)
17 or 18 metal oxide NPs	Cytotoxicity to <i>E. coli</i> and HaCaT cells	NP size NP agglomerate size	A combination of simple descriptors which reflect NPs' structure for the different levels of organization: from a single metal oxide molecule (i.e.	Sizochenko (2014)

(continued)

Table 7 (continued)

NMs used	End-point and biological system	Physico-chemical properties characterised	Calculated descriptors	References
229 NPs/cases from literature, 32 diverse chemistries	Toxicity against 1 out of 50 biological targets (b_1) with different complexities (algae, bacteria, cell lines, crustaceans, plants, fish, and others)	Some NPs reported in their bare forms, while other NPs were coated by different organic molecules, specifically using 12 different coating agents (s_c) NP size (L) based on experimental data	chemical structure) to a supramolecular ensemble of molecules (i.e. nanoparticle size). Liquid Drop model used for NP size. Distinction between interactions of surface versus internal atoms. Parameters to describe metal ion's affinity to biochemical ligands: covalent index (CI) and cation polarizing power (CPP). RF modeling obtained 6 significant descriptors for HaCaT keratinocytes and 7 descriptors for <i>E. coli</i> . Descriptors S1, rw, ρ are the same for both models (see Fig. 10)	Kleandrova (2014)
			4 different descriptors namely: <ul style="list-style-type: none"> • molar volume (V) • electronegativity (E) • polarizability (P) • NP size (L) Three descriptors were extracted from the public source Chemical Periodic Table (Hsu 2013) These were not predictive so new descriptors were calculated, for example:	(continued)

Table 7 (continued)

NMs used	End-point and biological system	Physico-chemical properties characterised	Calculated descriptors	References
300 NPs/cases retrieved from public source OChem	Anti-bacterial activity using range of assays in up to 34 bacterial strains NPs classified as active or inactive	Data from database • NP size • NP shape • NP surface coating (15 different) Molar volume (AMV), Electronegativity (AE) and Polarizability (AP), retrieved from website Chemical Periodic Table	<ul style="list-style-type: none"> • TEI(cj)rf a binary (classification) variable expressing the toxic effect of the NP used as reference • $\Delta\Delta E(bt)$ a perturbation term that describes changes in the electronegativity between new (output or final state) NM and the other used as reference, also depending on the biological targets (9 such descriptors in total) Mixed NPs calculated as sum of properties/number of atoms Chemical structures of coating agents computed: <ul style="list-style-type: none"> • spectral moments of bond adjacency matrix, weighted by properties such as standard bond distance, the atomic polar surface area and the atomic refractivity 	Speck-Planche (2015)
18 nano-metal oxides	Toxicity to human keratinocyte (HaCaT) cell line as a model for dermal exposure.	TEM and DLS for NP size	Calculated a set of 27 parameters quantitatively describing variability of NPs' structure—nano-descriptors including: 16 quantum-mechanical descriptors, i.e.	Gajewicz (2015)

(continued)

Table 7 (continued)

NMs used	End-point and biological system	Physico-chemical properties characterised	Calculated descriptors	References
Metal oxide nanoparticles	Comparison of NP toxicity to eukaryotic system (HaCaT) and the prokaryotic system (<i>E. coli</i>)		<ul style="list-style-type: none"> • Total energy • Electronic energy, • Core–core repulsion energy, • Solvent accessible-surface, • Energy of HOMO • Energy of LUMO • Chemical hardness, • Total softness, • HOMO-LUMO energy gap, • Electronic chemical potential, • Valance band, • Conduction band 11 image descriptors (derived from TEM images)—Area, Volume, Surface diameter, Volume/mass diameter, Volume/surface area, Aspect ratio X, Aspect ratio Y, Porosity X, Porosity Y, Circularity, Sphericity	Toropova (2015)
	Photo-induced cytotoxicity to bacteria <i>Escherichia coli</i>	NPs represented as molecular input-line entry system (SMILES)	Quasi-SMILES is the representation of data on molecular structure together with condition: presence or absence of photo-inducing. The presence of photo-inducing indicated by symbol ‘\^’ that is added at the end of traditional SMILES	(continued)

Table 7 (continued)

NMs used	End-point and biological system	Physico-chemical properties characterised	Calculated descriptors	References
105 surface modified Gold NPs utilising 67 different surface ligands	Dataset of cellular association of 84 Au NPs from Liu (2015a)	39 physico-chemical properties of the Au NPs (as synthesized and with serum) including TEM and DLS size characterization, zeta potential, absorbance spectrophotometry, and the amount of adsorbed serum protein on the NP surface obtained from the bicinchoninic acid (BCA) assay	Both linear and nonlinear (e-SVR) regressions used for cellular association of Au NPs. The regressions were coupled with sequential forward floating selection to identify suitable QSAR descriptors from the three descriptor sets, including 120 protein corona fingerprints, 19 of the nanoparticle properties, and a composite of the two sets	Liu (2015a)
Iron oxide core with different surface-modifying organic molecules	Cellular uptake	Unknown	11 found predictive in linear model 6 predictive in non-linear model	Liu (2015b)

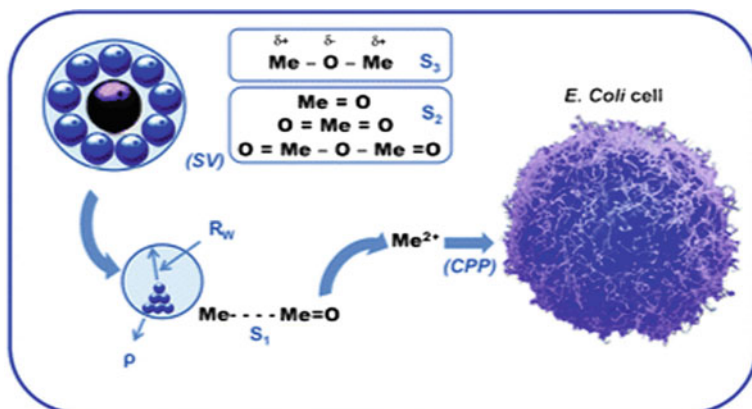


Fig. 10 Schematic representation of the mechanism of metal oxide nanoparticle toxicity for *E. coli* cells. From Sizochenko (2014). The important descriptors for *E. coli* cytotoxicity of the metal oxide NMs were determined to be: S_1 unbonded two-atomic fragments $[\text{Me}] \cdots [\text{Me}]$, which were encoded based on SiRMS-derived descriptors, encoding the distance where the potential reaches minimum at van der Waals interactions (7%); r_w , Wigner–Seitz radius (22%); ρ mass density (2%); CPP , cation polarizing power (30%); S_2 SiRMS-derived electronegativity aligned descriptor of oxides molecules—in a sense the acid–base property of oxides. This parameter increases with the number of oxygen atoms in a molecule (3%); S_3 tri-atomic fragments $[\text{Me}]-[\text{O}]-[\text{Me}]$, which were encoded by SiRMS-derived descriptors, encoding electronegativity (29%); and SV , the proportion of surface molecules to molecules in volume (7%). % is used to represent the absolute impacts for each descriptor

the various pathways involved. One potential starting point is to link with the emerging Adverse Outcome Pathways (AOP) approach, which aims to identify so-called molecular initiating events (MIE) and follow these through the subsequent effects to adverse outcomes and cellular, organismal, community and population levels. Initial conceptualization for nanotoxicology is already underway (Gerloff 2016), and a case study of liver toxicity proposed that the differences between NM and chemically-induced adversity were primarily related to differences in toxicokinetics and the nature of the initial Key Events in the AOP. NM reactivity was associated with the NM's potential to generate oxidative stress, determined as the ability of the NMs to exchange electrons with biological redox species in the cell (Gerloff 2016). The model was tested using the metal oxide libraries described in Table 5 (Zhang 2012) and was only partially accurate in predicting the capacity of metal oxide NMs to induce oxidative stress, since other metal oxide NMs induce similar effects through ion dissolution, illustrating the importance of relating QSAR properties to proposed MIEs (Gerloff 2016). Thus, teasing out the relative contributions of different NM physico-chemical parameters to specific MIEs, and grouping of NMs based on all properties that link to specific MIEs might be a useful way forward. One proposal of how to do such a mapping of the partial contributions of multiple parameters to a specific endpoint was proposed by Lynch et al., using principle components analysis which could then be complemented with factorial

analysis to determine contributions to the different modes of action (Lynch et al. 2014a). The commonality of NM physico-chemical parameters confirmed to contribute across the different modes of action shown in Table 2 shows the scale of the computational and experiment challenge, and highlights the need for much closer cooperation between the two approaches, and the need for involvement of modellers in co-designing experimental studies to ensure they provide the full spectrum of temporal and spatial datasets needed for computational and *in silico* approaches.

5 Conclusions

This chapter summarises the current state of the art in terms of mechanisms of NM toxicity, in light of the emerging understanding of the context-dependence of physico-chemical characteristics of NMs that may be specifically linked to their toxicity. In addition, the chapter summarises some of the challenges for defining and regulating NMs, and in particular for developing predictive read-across tools including QSARs/QPARs, which result from the fact that NMs are a highly diverse and highly dynamic group of materials. While there is growing consensus that many of the biological effects from manufactured or engineered NMs may not be all that different from ultrafine anthropogenic particles of similar compositions, it is clear that the variety of materials that can be engineered at the nanoscale, and the variety of forms (shapes, structures, morphologies, composites and hybrids etc.) that can be produced opens up significantly more challenges than arise from combustion-related particles.

In particular, the current state of the art in terms of NM libraries with accompanying physico-chemical characterisation and uptake or toxicity datasets available in the literature that can be used as the basis for development of QSARs/QPARs is presented. Building on this, some initial progress towards QSAR models was presented, and the challenges and ongoing disconnect between measured properties and those found to be predictive in QSARs, which are almost entirely calculated, often entirely independently from the measured datasets was discussed.

An ongoing challenge for the field is that access to the NMs themselves for additional experimentation/validation is generally limited outside the group that produced the NMs. Thus, it would be valuable for journals to require that NMs underpinning datasets be made available for sharing upon request or placement of samples in a repository, although there are significant cost challenges associated with this, as well as issues related to NMs ageing and evolution. However, it is unlikely that the costs would be more significant than those associated with protein synthesis or development of transgenic animal models, so approaches could be adopted from these fields alternatively. An alternative would be that full SOPs for production of the NMs be published (or included as part of the supplementary information) and detailed accompanying characterisation information (including

SOPs, instrument details and calibrations), and that authors are provided with facilities for cross-checking of outputs by the original NM-library developer, as part of the publication requirements for NMs datasets.

Acknowledgements This work was funded via the European Commission's 7th Framework Programme project "NanoMILE" (Contract N°. NMP4-LA-2013-310451). The authors acknowledge the NanoMILE consortium for constructive discussions.

References

- Ahmadi, T. S., Wang, Z. L., Green, T. C., Henglein, A., & El-Sayed, M. A. (1996). Shape-controlled synthesis of colloidal platinum nanoparticles. *Science*, *272*, 1924–1925.
- Albanese, A., Walkey, C. D., Olsen, J. B., Guo, H., Emili, A., & Chan, W. C. (2014). Secreted biomolecules alter the biological identity and cellular interactions of nanoparticles. *ACS Nano*, *8*, 5515–5526.
- Baun, A., Sørensen, S. N., Rasmussen, R. F., Hartmann, N. B., & Koch, C. B. (2008). Toxicity and bioaccumulation of xenobiotic organic compounds in the presence of aqueous suspensions of aggregates of nano-C(60). *Aquatic Toxicology*, *86*, 379–387.
- Bexiga, M. G., Varela, J. A., Wang, F., Fenaroli, F., Salvati, A., Lynch, I., et al. (2011). Cationic nanoparticles induce caspase 3-, 7- and 9-mediated cytotoxicity in a human astrocytoma cell line. *Nanotoxicology*, *5*, 557–567.
- Brus, L. (1994). Luminescence of silicon materials: Chains, sheets, nanocrystals, nanowires, microcrystals, and porous silicon. *Journal of Physical Chemistry*, *98*, 3575–3581.
- Burello, E., & Worth, A. P. (2011). A theoretical framework for predicting the oxidative stress potential of oxide nanoparticles. *Nanotoxicology*, *5*, 228–235.
- Bussy, C., Pinault, M., Cambedouzou, J., Landry, M. J., Jegou, P., Mayne-L'hermite, M., et al. (2012). Critical role of surface chemical modifications induced by length shortening on multi-walled carbon nanotubes-induced toxicity. *Particle and Fibre Toxicology*, *9*, 46.
- Chng, E. L., & Pumera, M. (2013). The toxicity of graphene oxides: Dependence on the oxidative methods used. *Chemistry*, *19*, 8227–8235.
- Chowdhury, I., Duch, M. C., Gits, C. C., Hersam, M. C., & Walker, S. L. (2012). Impact of synthesis methods on the transport of single walled carbon nanotubes in the aquatic environment. *Environmental Science and Technology*, *46*, 11752–11760.
- Clemments, A. M., Botella, P., & Landry, C. C. (2015). Protein adsorption from biofluids on silica nanoparticles: Corona analysis as a function of particle diameter and porosity. *ACS Applied Materials & Interfaces*, *7*, 21682–21689.
- Dawson, K. A., Linse, S., & Lynch, I. (2007). Water as a mediator of protein-nanoparticle interactions: Entropy driven protein binding as a paradigm for protein therapeutics in the Biopharma industry? *E-nano Newsletter [Online]*, 23–34.
- Deng, Z. J., Liang, M., Monteiro, M., Toth, I., & Minchin, R. F. (2011). Nanoparticle-induced unfolding of fibrinogen promotes Mac-1 receptor activation and inflammation. *Nature Nanotechnology*, *6*, 39–44.
- Deng, Z. J., Liang, M., Toth, I., Monteiro, M., & Minchin, R. F. (2013). Plasma protein binding of positively and negatively charged polymer-coated gold nanoparticles elicits different biological responses. *Nanotoxicology*, *7*, 314–322.
- Fadeel, B., Feliu, N., Vogt, C., Abdelmonem, A. M., & Parak, W. J. (2013). Bridge over troubled waters: understanding the synthetic and biological identities of engineered nanomaterials. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology*, *5*, 111–129.

- Gajewicz, A., Schaublin, N., Rasulev, B., Hussain, S., Leszczynska, D., Puzyn, T., et al. (2015). Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles: Hints from nano-QSAR studies. *Nanotoxicology*, *9*, 313–325.
- George, S., Lin, S. J., Jo, Z. X., Thomas, C. R., Li, L. J., Mecklenburg, M., et al. (2012). Surface defects on plate-shaped silver nanoparticles contribute to its hazard potential in a fish gill cell line and zebrafish embryos. *ACS Nano*, *6*, 3745–3759.
- Gerloff, K., Landesmann, B., Worth, A., Munn, S., Palosaari, T., & Whelan, M. (2016). The adverse outcome pathway approach in nanotoxicology. *Computational Toxicology*.
- Harper, S. L., Carriere, J. L., Miller, J. M., Hutchinson, J. E., Maddux, B. L. S., & Tanguay, R. L. (2011). Systematic evaluation of nanomaterial toxicity: Utility of standardised materials and rapid assays. *ACS Nano*, *5*, 4688–4697.
- Hasselölv, M., & Kaegi, R. (2009). Analysis and characterization of manufactured nanoparticles in aquatic environments. In J. R. Lead & E. Smith (Eds.), *Environmental and human health impacts of nanotechnology*.
- Ho, C.-M., Yau, S. K.-W., Lok, C.-N., So, M.-H., & Che, C.-M. (2010). Oxidative dissolution of silver nanoparticles by biologically relevant oxidants: A kinetic and mechanistic study. *Chemistry—An Asian Journal*, *5*, 285–293.
- HSU, D. D. (2013). Chemical periodic table. <http://www.chemicool.com/>.
- Ivask, A., Suarez, E., Patel, T., Boren, D., Ji, Z. X., Holden, P., et al. (2012). Genome-wide bacterial toxicity screening uncovers the mechanisms of toxicity of a cationic polystyrene nanomaterial. *Environmental Science and Technology*, *46*, 2398–2405.
- Kleandrova, V. V., Luan, F., González-Díaz, H., Ruso, J. M., Speck-Planche, A., & Cordeiro, N. D. S. (2014). Computational tool for risk assessment of nanomaterials: Novel QSTR-perturbation model for simultaneous prediction of ecotoxicity and cytotoxicity of uncoated and coated nanoparticles under multiple experimental conditions. *Environmental Science and Technology*, *48*, 14686–14694.
- Klein, J. (2007). Probing the interactions of proteins and nanoparticles. *PNAS*, *104*, 2029–2030.
- Lee, K. J., Browning, L. M., Nallathamby, P. D., & Xu, X. H. (2013). Study of charge-dependent transport and toxicity of Peptide-functionalized silver nanoparticles using zebrafish embryos and single nanoparticle plasmonic spectroscopy. *Chemical Research in Toxicology*, *26*, 904–917.
- Li, R., Ji, Z., Chang, C. H., Dunphy, D. R., Cai, X., Meng, H., et al. (2014). Surface interactions with compartmentalized cellular phosphates explain rare earth oxide nanoparticle hazard and provide opportunities for safer design. *ACS Nano*, *8*, 1771–1783.
- Li, Y., Zhang, W., Niu, J., & Chen, Y. (2013). Surface coating-dependent dissolution, aggregation, and ROS generation of silver nanoparticles under different irradiation conditions. *Environmental Science and Technology*, *47*, 10293–10301.
- Lin, S., Zhao, Y., Ji, Z., et al. (2013). Zebrafish high-throughput screening to study the impact of dissolvable metal oxide nanoparticles on the hatching enzyme, ZHE1. *Small (Weinheim an der Bergstrasse, Germany)* *9*, 1776–1785. doi:10.1002/sml.201202128.
- Linsinger, T., Roebben, G., Gilliland, D., Calzolari, L., Rossi, F., Gibson, N., et al. (2012). Requirements on measurements for the implementation of the European Commission definition of the term “nanomaterial”. *Report EUR 25404 EN*.
- Liu, J., Aruguete, D. M., Jinschek, J. R., Rimstidt, J. D., & Hochella, Jr., M. F. (2008). The non-oxidative dissolution of galena nanocrystals: Insights into mineral dissolution rates as a function of grain size, shape, and aggregation state. *Geochimica et Cosmochimica Acta*, *72*, 5984–5996.
- Liu, J., von der Kammer, F., Zhang, B., Legros, S., & Hofmann, T. (2013a). Combining spatially resolved hydrochemical data with in-vitro nanoparticle stability testing: Assessing environmental behavior of functionalized gold nanoparticles on a continental scale. *Environment International*, *59*, 53–62.
- Liu, R., Jiang, W., Walkey, C. D., Chan, W. C., & Cohen, Y. (2015a). Prediction of nanoparticles-cell association based on corona proteins and physicochemical properties. *Nanoscale*, *7*, 9664–9675.

- Liu, R., Rallo, R., Bilal, M., & Cohen, Y. (2015b). Quantitative structure-activity relationships for cellular uptake of surface-modified nanoparticles. *Combinatorial Chemistry & High Throughput Screening*, *18*, 365–375.
- Liu, R., Rallo, R., George, S., Ji, Z., Nair, S., Nel, A. E., et al. (2011). Classification NanoSAR development for cytotoxicity of metal oxide nanoparticles. *Small*, *7*, 1118–1126.
- Liu, X., Chen, G., Keller, A. A., & Su, C. (2013b). Effects of dominant material properties on the stability and transport of TiO₂ nanoparticles and carbon nanotubes in aquatic environments: From synthesis to fate. *Environmental Science: Processes Impacts*, *15*, 169–189.
- Lövestam, G., Rauscher, H., Roebben, G., Sokull Klüttgen, B., Gibson, N., Putaud, J.-P., et al. (2010). Considerations on a definition of nanomaterial for regulatory purposes.
- Lowry, G. V., Gregory, K. B., Apte, S. C., & Lead, J. R. (2012). Transformations of nanomaterials in the environment. *Environmental Science and Technology*, *46*, 6893–6899.
- Lowry, G. V., Hill, R. J., Harper, S., Rawle, A. F., Hendren, C. O., Klaessig, F., et al. (2016). Guidance to improve the scientific value of zeta-potential measurements in nanoEHS. *Environmental Science: Nano*, *3*, 953–965.
- Lynch, I. (2007). Are there generic mechanisms governing interactions between nanoparticles and cells? Epitope mapping the outer layer of the protein-material interface. *Physica A—Statistical Mechanics and its Applications*, *373*, 511–520.
- Lynch, I., Dawson, K. A., Lead, J. R., & Valsami-Jones, E. (2014a). Macromolecular coronas and their importance in nanotoxicology and nanocotoxicology. In J. R. Lead & E. Valsami-Jones (Eds.), *Nanoscience and the environment*.
- Lynch, I., Weiss, C., & Valsami-Jones, E. (2014b). A strategy for grouping of nanomaterials based on key physico-chemical descriptors as a basis for safer-by-design NMs. *Nano Today*, *9*, 266–270.
- Melagraki, G., & Afantitis, A. (2014). Enalos InSilicoNano Platform: An online decision support tool for the design and virtual screening of nanoparticles. *RSC Advances*, *4*, 50713–50725.
- Melagraki, G., & Afantitis, A. (2015). A risk assessment tool for the virtual screening of metal oxide nanoparticles through Enalos InSilicoNano Platform. *Current Topics in Medicinal Chemistry*, *15*, 1827–1836.
- Meng, H., Xia, T., George, S., & Nel, A. E. (2009). A predictive toxicological paradigm for the safety assessment of nanomaterials. *ACS Nano*, *3*, 1620–1627.
- Merhi, M., Dombu, C. Y., Brient, A., Chang, J., Platel, A., le Curieux, F., et al. (2012). Study of serum interaction with a cationic nanoparticle: Implications for in vitro endocytosis, cytotoxicity and genotoxicity. *International Journal of Pharmaceutics*, *423*, 37–44.
- Misra, S. K., Dybowska, A., Berhanu, D., Luoma, S. N., & Valsami-Jones, E. (2012). The complexity of nanoparticle dissolution and its importance in nanotoxicological studies. *Science of the Total Environment*, *438*, 225–232.
- Nap, R. J., & Szeleifer, I. (2013). How to optimize binding of coated nanoparticles: Coupling of physical interactions. *Molecular Organization and Chemical State Biomaterial Science*, *1*, 814–823.
- Nel, A. E., Parak, W. J., Chan, W. C., Xia, T., Hersam, M. C., Brinker, C. J., et al. (2015). Where are we heading in nanotechnology environmental health and safety and materials characterization? *ACS Nano*, *9*, 5627–5630.
- Pagnout, C., Jomini, S., Dadhwal, M., Caillet, C., Thomas, F., & Bauda, P. (2012). Role of electrostatic interactions in the toxicity of titanium dioxide nanoparticles toward *Escherichia coli*. *Colloids and Surfaces B: Biointerfaces*, *92*, 315–321.
- Park, E.-J., Yi, J., Kim, Y., Choi, K., & Park, K. (2010). Silver nanoparticles induce cytotoxicity by a Trojan-horse type mechanism. *Toxicology in Vitro*, *24*, 872–878.
- Petersen, L. K., Chavez-Santoscoy, A. V., & Narasimhan, B. (2012). Combinatorial synthesis of and high-throughput protein release from polymer film and nanoparticle libraries. *Journal of Visualized Experiments*, *67*, 3882.
- Pokhrel, S., Nel, A. E., & Mädler, L. (2012). Custom-designed nanomaterial libraries for testing metal oxide toxicity. *Accounts of Chemical Research*, *46*, 632–641.

- Puzyn, T., Rasulev, B., Gajewicz, A., Hu, X., Dasari, T. P., Michalkova, A., et al. (2011). Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nature Nanotechnology*, *6*, 175–178.
- Salvati, A., Aberg, C., dos Santos, T., Varela, J., Pinto, P., Lynch, I., et al. (2011). Experimental and theoretical comparison of intracellular import of polymeric nanoparticles and small molecules: Toward models of uptake kinetics. *Nanomedicine*, *7*, 818–826.
- Sau, T. K., & Murphy, C. J. (2004). Room temperature, high-yield synthesis of multiple shapes of gold nanoparticles in aqueous solution. *Journal of the American Chemical Society*, *126*, 8648–8649.
- Sau, T. K., Urban, A. S., Dondapati, S. K., Fedoruk, M., Horton, M. R., Rogach, A. L., et al. (2009). Controlling loading and optical properties of gold nanoparticles on liposome membranes. *Colloids and Surfaces A—Physicochemical and Engineering Aspects*, *342*, 92–96.
- Shaw, S. Y., Westly, E. C., Pittet, M. J., Subramanian, A., Schreiber, S. L., & Weissleder, R. (2008). Perturbational profiling of nanomaterial biologic activity. *Proceedings of the National Academy of Sciences U.S.A.*, *105*, 7387–7392.
- Sizochenko, N., Rasulev, B., Gajewicz, A., Kuz'Min, V., Puzyn, T., & Leszczynski, J. (2014). From basic physics to mechanisms of toxicity: The “liquid drop” approach applied to develop predictive classification models for toxicity of metal oxide nanoparticles. *Nanoscale*, *6*, 13986–13993.
- Soler, M. A. G., Lima, E. C. D., da Silva, S. W., Melo, T. F. O., Pimenta, A. C. M., Sinnecker, J. P., et al. (2007). Aging investigation of cobalt ferrite nanoparticles in low pH magnetic fluid. *Langmuir*, *23*, 9611–9617.
- Speck-Planche, A., Kleandrova, V. V., Luan, F., & Cordeiro, M. N. D. S. (2015). Computational modeling in nanomedicine: Prediction of multiple antibacterial profiles of nanoparticles using a quantitative structure-activity relationship perturbation model. *Nanomedicine*, *10*, 193–204.
- Stamm, H. (2011). Risk factors: Nanomaterials should be defined. *Nature*, *476*, 399.
- Stefaniak, A. B., Hackley, V. A., Roebben, G., Ehara, K., Hankin, S., Postek, M. T., et al. (2013). Nanoscale reference materials for environmental, health and safety measurements: Needs, gaps and opportunities. *Nanotoxicology*, *7*, 1325–1337.
- Studer, A. M., Limbach, L. K., van Duc, L., Krumeich, F., Athanassiou, E. K., Gerber, L. C., et al. (2010). Nanoparticle cytotoxicity depends on intracellular solubility: Comparison of stabilized copper metal and degradable copper oxide nanoparticles. *Toxicology Letters*, *197*, 169–174.
- Sund, J., Alenius, H., Vippola, M., Savolainen, K., & Puustinen, A. (2011). Proteomic characterization of engineered nanomaterial–protein interactions in relation to surface reactivity. *ACS Nano*, *5*, 4300–4309.
- Teoh, W. Y., Amal, R., & Madler, L. (2010). Flame spray pyrolysis: An enabling technology for nanoparticles design and fabrication. *Nanoscale*, *2*, 1324–1347.
- Toropova, A. P., Toropov, A. A., Rallo, R., Leszczynska, D., & Leszczynski, J. (2015). Optimal descriptor as a translator of eclectic data into prediction of cytotoxicity for metal oxide nanoparticles under different conditions. *Ecotoxicology and Environmental Safety*, *112*, 39–45.
- Tsuzuki, T. (2009). Commercial scale production of inorganic nanoparticles. *International Journal of Nanotechnology (IJNT)*, *6*. doi:10.1504/IJNT.2009.024647.
- Vácha, R., Martínez-Veracoechea, F. J., & Frenkel, D. (2011). Receptor-mediated endocytosis of nanoparticles of various shapes. *Nano Letters*, *11*, 5391–5395.
- Valsami-Jones, E., & Lynch, I. (2015). How safe are nanomaterials? *Science*, *350*, 388–389.
- von der Kammer, F., Ferguson, P. L., Holden, P. A., Masion, A., Rogers, K. R., Klaine, S. J., et al. (2012). Analysis of engineered nanomaterials in complex matrices (environment and biota): General considerations and conceptual case studies. *Environmental Toxicology and Chemistry*, *31*, 32–49.
- Walczyk, D., Bombelli, F. B., Monopoli, M. P., Lynch, I., & Dawson, K. A. (2010). What the cell “sees” in bionanoscience. *Journal of the American Chemical Society*, *132*, 5761–5768.
- Walkey, C. D., Olsen, J. B., Song, F., Liu, R., Guo, H., Olsen, D. W., et al. (2014). Protein corona fingerprinting predicts the cellular interaction of gold and silver nanoparticles. *ACS Nano*, *8*, 2439–2455.

- Wang, X., Duch, M. C., Mansukhani, N., Ji, Z., Liao, Y. P., Wang, M., et al. (2015). Use of a pro-fibrogenic mechanism-based predictive toxicological approach for tiered testing and decision analysis of carbonaceous nanomaterials. *ACS Nano*, *9*, 3032–3043.
- Weissleder, J., Kelly, K., Sun, E. Y., Shtatland, T., & Josephson, L. (2005). Cell-specific targeting of nanoparticles by multivalent attachment of small molecules. *Nature Biotechnology*, *23*, 1418–1423.
- Wells, D. M., Rossi, G., Ferrando, R., & Palmer, R. E. (2015). Metastability of the atomic structures of size-selected gold nanoparticles. *Nanoscale*, *7*, 6498–6503.
- Xia, T., Meng, H., George, S., Zhang, H., Wang, X., Ji, Z., et al. (2012). Strategy for toxicity screening of nanomaterial. *Material Matters*. <http://www.sigmaaldrich.com/technical-documents/articles/materials-science/strategy-for-toxicity-screening-of-nanomaterials.html>.
- Xia, X.-R., Monteiro-Riviere, N. A., & Riviere, J. E. (2010). An index for characterization of nanomaterials in biological systems. *Nature Nanotechnology*, *5*, 671–675.
- Yang, S. T., Liu, Y., Wang, Y. W., & Cao, A. (2013). Biosafety and bioapplication of nanomaterials by designing protein–nanoparticle interactions. *Small*, *9*, 1635–1653.
- Zhang, H., Dunphy, D. R., Jiang, X., Meng, H., Sun, B., Tam, D., et al. (2012a). Processing pathway dependence of amorphous silica nanoparticle toxicity: Colloidal vs pyrolytic. *JACS*, *134*, 15790–15804.
- Zhang, H., Ji, Z., Xia, T., Meng, H., Low-Kam, C., Liu, R., et al. (2012b). Use of metal oxide nanoparticle band gap to develop a predictive paradigm for oxidative stress and acute pulmonary inflammation. *ACS Nano*, *6*, 4349–4368.
- Zheng, H., Mortensen, L. J., & Delouise, L. A. (2013). Thiol antioxidant-functionalized CdSe/ZnS quantum dots: Synthesis, characterization, cytotoxicity. *Journal of Biomedical Nanotechnology*, *9*, 382–392.

In Silico Approaches for the Prediction of In Vivo Biotransformation Rates

Ester Papa, Jon A. Arnot, Alessandro Sangion and Paola Gramatica

Abstract The assessment of chemical bioaccumulation is a required procedure under several regulatory frameworks. However, since the experimental quantification of bioaccumulation and related metrics (such as the Bioconcentration Factor, BCF) is resource intensive (money, animals) and time consuming, several computational approaches have been proposed as an alternative. Most bioaccumulation model estimates based on the octanol water partition coefficient (K_{OW}) alone can be inaccurate, if they do not take into account additional processes that influence chemical partitioning, chemical uptake and elimination rates. In particular, the biotransformation rate constant (k_B) can play a significant role in mitigating the bioaccumulation potential of hydrophobic chemicals. Bioaccumulation model (e.g., BCF) estimates can be refined when experimental or predicted k_B values are available. The aim of this chapter is to illustrate the development and the application of in silico models for in vivo biotransformation rates, for the cost-effective estimation of k_B for screening assessment. The chapter includes several examples of

E. Papa (✉) · A. Sangion · P. Gramatica
Department of Theoretical and Applied Sciences, QSAR Research
Unit in Environmental Chemistry and Ecotoxicology,
University of Insubria, Varese, Italy
e-mail: ester.papa@uninsubria.it

A. Sangion
e-mail: alessandro.sangion@uninsubria.it

P. Gramatica
e-mail: paola.gramatica@uninsubria.it

J.A. Arnot
ARC Arnot Research & Consulting, Toronto, ON, Canada
e-mail: jon@arnotresearch.com

J.A. Arnot
Department of Physical and Environmental Science, University of Toronto Scarborough,
Toronto, ON, Canada

J.A. Arnot
Department of Pharmacology and Toxicology,
University of Toronto, Toronto, ON, Canada

quantitative structure-activity relationships (QSARs), which predict k_B or the associated half-life from the chemical structure. Furthermore, the chapter describes the complementary role of in vitro biotransformation rate estimation and the subsequent in vitro-to-in vivo extrapolation (IVIVE) calculations for refining bioaccumulation model predictions.

Keywords Biotransformation prediction • Bioaccumulation refinement
In silico • QSARINS • IVIVE • Metabolism

1 Introduction

The correct identification and quantification of properties like Persistence, Bioaccumulation and Toxicity of chemicals represents a primary issue for the assessment of chemical hazards and potential risks posed to humans and the environment (European Union, REACH regulation 2006; Cowan-Ellsberry et al. 2008; European Chemicals Agency 2008; Lillicrap et al. 2016; UNEP 2016). The investigation of the potential effects due to the accumulation of substances in humans and animals has been a challenging topic in the last few decades (Arnot and Gobas 2006; Kim et al. 2016; Mackay et al. 2016). The assessment of bioaccumulation potential is a required procedure under several regulatory frameworks, and different bioaccumulation assessment metrics are available, such as the Bioconcentration Factor (BCF) and the Bioaccumulation Factor (BAF). The BCF is measured under controlled laboratory experiments in which the test organism (typically fish) is exposed to chemical in the water only (OECD 2012). The BAF is measured in the environment in which the organisms are exposed to chemical from the surrounding environment and their diet (Burkhard et al. 2012). A main limitation in bioaccumulation studies is the large cost of laboratory experiments and field studies necessary to quantify the bioaccumulation metrics and related process parameters. A study by Arnot and Gobas (2006) highlighted that less than 5% of commercial organic chemicals have measured BCF or BAF data for fish. This suggests the importance of modelling approaches, such as those based on quantitative structure-activity relationships (QSARs) or mechanistic mass balance models, to predict the missing information and gain additional knowledge in this area of research. In addition, the use of QSAR models is suggested under several regulatory frameworks as a cost effective solution to be adopted in the absence of experimental data, which can support and integrate hazard assessment procedures (e.g., European Union, REACH regulation 2006; European Union, Cosmetic Regulation 2009; European Union, Biocidal Products Regulation 2012). Several QSAR approaches have been proposed for the prediction of bioaccumulation related parameters mainly based on physical-chemical properties, in particular those related to chemical hydrophobicity such as water solubility and the octanol-water partition coefficient (K_{OW}). However, estimates based on chemical properties alone can be inaccurate since they do not take into account bioaccumulation processes such as

biotransformation (also commonly referred to as metabolism) (Veith et al. 1979; Mackay 1982; Cowan-Ellsberry et al. 2008; Arnot et al. 2009; Segner 2015).

The first-order, whole body, primary biotransformation rate constant (k_B) and corresponding half-life (HL) play an important role in mitigating the bioaccumulation potential of hydrophobic chemicals in aquatic organisms. For example, when the BCF estimation is refined with in vitro experimental or QSAR predicted k_B values, the calculated BCFs are in better agreement with measured BCFs (Cowan-Ellsberry et al. 2008; Laue et al. 2014; Segner 2015). Biotransformation was highlighted in the 1980s as a determinant factor for the refinement of BAF related estimations in fish (Lech and Bend 1980) and continues to be indicated as a key element for bioaccumulation science (Burkhard et al. 2012). The influence of biotransformation on BAF related processes has been investigated by several experimental and modelling approaches (e.g., Burkhard 2003; Arnot et al. 2008a; Barber 2008; Kim et al. 2016; Mackay et al. 2016). Moreover, the biotransformation process has been studied at different levels of complexity and biological organization (e.g., cellular or whole body), and by different methods, which include experimental, in vitro and in vivo studies and in silico quantification of biotransformation kinetic parameters, metabolic pathways, and mechanisms (Wilk-Zasadna et al. 2015). Several in silico simulators are currently available for the prediction of metabolic pathways, metabolites, and/or of molecular sites potentially susceptible to biotransformation, such as METEORTM (Lhasa Ltd.), METATM (MultiCASE Inc.), PASS, (Borodina et al. 2003) CATABOL, and TIMES (Dimitrov et al. 2011; Mekenyan 2012). More examples have been recently reviewed by Peach and colleagues (Peach et al. 2012).

Other studies (Long and Walker 2003; Pirovano et al. 2012, 2015) describe quantitative models for the prediction, from the molecular structure of chemicals, of kinetic parameters (i.e., intrinsic clearance (CL_{INT}), Michaelis-Menten constant (K_M) and maximum velocity of the reaction (V_{MAX})) for different enzymatic reactions. In vitro-to-in vivo Extrapolation (IVIVE) methods for extrapolating in vitro metabolism rate data to liver clearance in mammals have been developed and applied in the pharmaceutical industry for decades (Wilkinson and Shand 1975). More recently, these types of in vitro assays (commonly using hepatocytes and microsomal or S9 liver fractions) and IVIVE approaches have been developed and applied to estimate hepatic clearance and whole body k_B values and refine model calculated BCFs in fish (Nichols et al. 2006, 2007, 2013; Cowan-Ellsberry et al. 2008). Methods to estimate empirically-based in vivo k_B values from laboratory experimental testing data, such as BCFs and other toxicokinetic studies, have been developed (de Wolf et al. 1993; Van der Linde et al. 2001; Arnot et al. 2008a), and from these methods k_B databases have been created (Arnot et al. 2008b). From these databases of in vivo whole body k_B estimates various QSAR approaches have been applied to generate and evaluate models to predict k_B from chemical structure (Arnot et al. 2009, 2014; Brown et al. 2012; Kuo and Di Toro 2013; Papa et al. 2014, 2016).

The aim of this chapter is to illustrate the challenging role of developing alternative methods to animal testing, and in particular of in silico models, for the

cost-effective estimation of k_B . To this end, a short overview of biotransformation processes and related parameters is reported in Sect. 2. Strengths and limitations of *in silico* and *in vitro* methods are commented in Sects. 3 and 4. In particular, Sect. 3 describes several examples of models based on QSARs, which are available to predict k_B or the associated half-life from the chemical structure. Section 4 describes the general approach for using *in vitro* biotransformation assays and IVIVE models to estimate hepatic clearance *in vivo* and related *in silico* models.

Although the intention of this chapter is not to be a comprehensive review of all available datasets and models for predicting *in vivo* biotransformation rates, we hope that it can provide leads to readers interested in learning more about this active area of scientific research.

2 Biotransformation Processes

The fate of chemicals in the environment is heavily determined by degradation processes that transform the original structure of a compound (i.e., the parent compound) into another chemical structure. Degradations may be the result of abiotic reactions, such as photolysis or hydrolysis after direct interaction with sunlight and water, respectively, or may be operated by organisms (i.e., biotransformation) (Fig. 1).

In particular, microbial biotransformation may lead to the complete mineralization of chemicals (i.e., transformation to carbon dioxide and water), also known as ultimate biodegradation, or to intermediate compounds within different metabolic pathways.

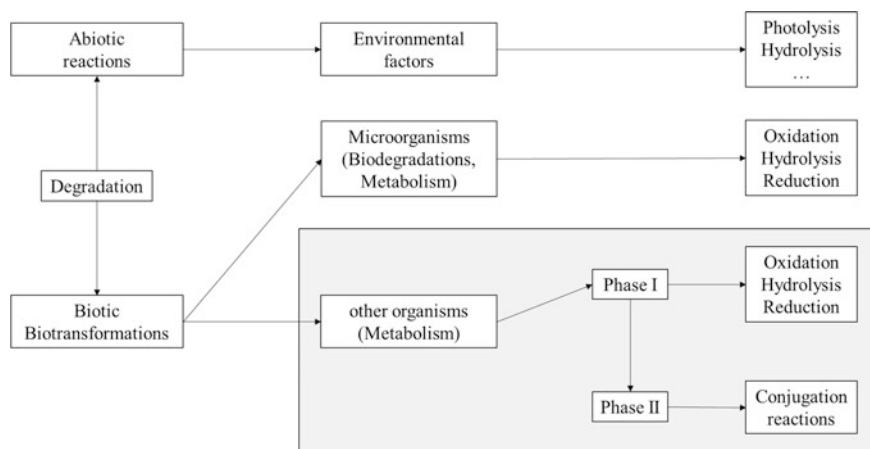


Fig. 1 Scheme of the main degradation reactions at abiotic and biotic level. Models commented in Sect. 3 are referred to biotic biotransformation operated by fish and mammals (i.e., within the grey shaded rectangle)

Biotransformation by higher organisms (e.g., fish or mammals) includes all the metabolic reactions which involve normal body constituents (e.g., lipids, proteins and carbohydrates) or xenobiotics: “*a man-made chemical or material not produced in nature and not normally considered a constituent component of a specified biological system*” (van Leeuwen and Vermeire 2007). These biotransformation processes generally occur in a two-phase series of reactions which can occur in multiple tissues but for most xenobiotics primarily in the liver (Weisbrod et al. 2009; Seviour et al. 2012; Walker et al. 2012; Mekenyan et al. 2012). During Phase I reactions, hydrophobic chemicals are typically functionalised to generate more polar (water soluble) metabolites. During Phase II reactions compounds are conjugated to large molecules to further increase polarity and generate more water soluble metabolites, which may be more easily excreted from the organism (Walker et al. 2012).

Phase I enzymes are located mainly in the endoplasmic reticulum. In particular, enzymes of the Cytochrome P450 system (CYP) are responsible for Phase I reactions (Seviour et al. 2012; Walker et al. 2012; Mekenyan et al. 2012). The principal enzymes in Phase II conjugation reactions are Glucuronyl transferases (UGT), Glutathione-S-transferases (GST) and Sulfotransferases (SULT) which are located on the endoplasmic reticulum and in the cytosol (Walker et al. 2012).

Biotransformation reactions seek to detoxify an organism by removing xenobiotics from the body; however, in some cases biotransformation may increase toxicity, leading to metabolites that are more toxic than the respective parent compound. Some examples of bioactivation are the generation of toxic epoxides, such as in the conversion of aldrin to dieldrin and of heptachlor to heptachlor epoxide, as well as the formation of DNA-binding epoxides such as in the bio-activation of benzo-[a]-pyrene (Sijm et al. 2007). Therefore, the identification of metabolites, and metabolic pathways, and the quantification of biotransformation rate parameters are critical steps in the determination of the possible toxic profile of chemicals (Arnot and Gobas 2003; Arnot et al. 2008b; Pirovano et al. 2016). Several approaches are available for the quantification of biotransformation using different testing strategies and models (Sijm et al. 2007; Weisbrod et al. 2009).

Biotransformation rates can be measured in vitro and in vivo by quantifying the rate of formation of a biotransformation product from a parent compound or by determining the rate of chemical loss (i.e., substrate depletion) in the defined system. For the former method, the metabolites of a particular chemical must be known. Measured data can then be used to develop in silico approaches for predicting rates and pathways (Borodina et al. 2003; Long and Walker 2003; Arnot et al. 2009, 2014; Dimitrov et al. 2011; Brown et al. 2012; Mekenyan et al. 2012; Papa et al. 2014; Pirovano et al. 2016). A list of the main parameters available to quantify different aspects of the biotransformation process (which will be commented on in the following sections of this chapter) is reported in Table 1.

Table 1 List of the main biotransformation related parameters used in QSAR and IVIVE approaches. See e.g., Nichols et al. (2013) for more details

Parameter	Units (e.g.,)	Definition	Derivation (e.g.,)
K_M	pmol/mL	Michaelis-Menten (affinity) constant—substrate concentration resulting in half-maximal activity	From experimental data
V_{MAX}	pmol/h/mg-protein	Maximum rate of the reaction	From experimental data
$k_{IN\ VITRO,\ INT}$	1/h	In vitro depletion rate constant	From experimental data
$CL_{IN\ VITRO,\ INT}$	ml/h/mg protein or mg/h/10 ⁶ cells	In vitro intrinsic clearance	V_{MAX}/K_M or $k_{IN\ VITRO,\ INT} \times C_{S9}$ or $k_{IN\ VITRO,\ INT} \times C_{HEP}$
$CL_{IN\ VIVO,\ INT}$	L/d/kg	In vivo intrinsic clearance	$CL_{IN\ VITRO,\ INT} L_{HEP} L_{FBW} \times 24$ or $CL_{IN\ VITRO,\ INT} L_{S9} L_{FBW} \times 24$
CL_H	L/d/kg	Hepatic clearance	$Q_H f_U CL_{IN\ VIVO,\ INT} / (Q_H + f_U CL_{IN\ VIVO,\ INT})$
k_T	1/d	Whole body, total elimination rate constant	Empirically-derived from in vivo experiment
k_B	1/d	Whole body, primary biotransformation rate constant	Empirically-derived from in vivo experiment or from in vivo experimental data and models (e.g., Arnot et al. 2008) or from IVIVE methods, e.g., $k_B = CL_H / V_D$
HL_T	d	Whole body, total elimination half-life	$\ln 2 / k_T$
HL_B	d	Whole body, primary biotransformation half-life	$\ln 2 / k_B$

C_{S9} —S9 protein concentration; C_{HEP} —hepatocyte concentration; L_{HEP} —hepatocellularity; L_{S9} —liver S9 protein concentration; L_{FBW} —fraction liver weight to whole body weight; Q_H —hepatic blood flow; f_U —hepatic clearance binding term; V_D —apparent volume of distribution referenced to blood (L/kg)

3 QSAR Prediction of in Vivo Biotransformation Rates and Half-Lives

As was mentioned in Sect. 1, few QSARs are currently available to predict k_B and the corresponding HLs (Arnot et al. 2009, 2014; Brown et al. 2012; Kuo and Di Toro 2013; Papa et al. 2014). These models are mathematical expressions of the quantitative relationships between the molecular structure of substances and the inferred in vivo biotransformation rates in fish and mammals. These QSARs are described in the following Sects. 3.1 and 3.2 where we first discuss the k_B data used for their development.

3.1 Models for Predicting Biotransformation Rates of Organic Chemicals in Fish

Prior to 2008 in vivo k_B estimates for fish were limited to about 30 chemicals (Arnot et al. 2008a). A kinetic mass balance method to obtain in vivo estimates of k_B from laboratory testing data was developed (Arnot et al. 2008a). Laboratory tests provide BCFs and total elimination rate constants k_T and half-lives. The general approach for the in vivo k_B estimation method is to use a model to predict rates of chemical elimination in fish for major routes of chemical elimination *other than biotransformation* and subtract the sum of these predicted rates (k_x) from the measured total elimination as (Arnot et al. 2008a):

$$k_B = k_T - k_x$$

This method includes an uncertainty analysis to propagate the uncertainty in the measured and predicted data (e.g., k_T and k_x , respectively) into the k_B estimate. The estimation method was corroborated with other available k_B estimates and subsequently applied to a large dataset of critically evaluated laboratory bioaccumulation data in fish to develop a database of approximately 700 discrete organic chemicals (Arnot et al. 2008b). Experimental data were measured in freshwater fish of different species mainly represented by *Oncorhynchus mykiss*, *Pimephales promelas*, *Cyprinus carpio* and *Poecilia reticulata*. Temperature and body size are known to influence toxicokinetics in fish (Peters 1983; Hendriks et al. 2001). In order to address some of the variability in the experimental data, k_B estimates were standardised to a mass and temperature specific rate constant (referred to as $k_{B,N}$) for a 10 g fish at 15 °C (median values from the database).

In addition, to characterize the uncertainty and assumed reliability of the k_B estimates, the empirically derived $k_{B,NS}$ were assigned to six categories on the basis of data confidence (very high, high, good, moderate, low, and uncertain) (Arnot et al. 2008b). Finally, $k_{B,N}$ values were converted to the respective primary biotransformation half-lives (HL_N) standardized for mass and temperature as:

$$HL_N = \frac{\ln 2}{k_{B,N}}$$

The $k_{B,N}$ (HL_N) dataset is included in the U.S. Environmental Protection Agency's Estimation Program Interface EPI Suite™ package Ver.4.1 (<https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface>), in the OECD Toolbox (<http://www.oecd.org/chemicalsafety/risk-assessment/theoecdqsar-toolbox.htm>) and is available on-line at <http://www.arnotresearch.com>.

This k_B (HL_N) dataset is highly heterogeneous including among others halogenated organics (polychlorinated biphenyls, dioxins and furans), aliphatic and

aromatic hydrocarbons, amines, imides, alcohols, phenols, ethers, ketones and esters etc..., as well as fluorinated compounds and siloxanes. Due to a general lack of measured bioaccumulation data in fish for highly dissociated ionisable organic chemicals (IOCs) and the uncertainty in estimating rate constants (k_X) for IOCs, the database does not include many IOCs. Nonetheless, the available data indicate that k_B values in fish span approximately 6 orders of magnitude. The database is for discrete organic chemicals only.

3.1.1 EPI Suite BCFBAF-QSAR

The first QSAR for the prediction of k_B in fish from the chemical structure of a heterogeneous set of 634 organic chemicals was developed by Arnot and colleagues in 2009 (Arnot et al. 2009).

A multiple linear regression model (which is included in the EPI Suite™ software and is here named EPI-QSAR) was derived by the method of ordinary least squares using 59 descriptors of which 57 are molecular fragments, in addition to K_{OW} and molar mass (MW) (Arnot et al. 2009).

The EPI-QSAR was developed after splitting the available data into training and the prediction set used for external validation on the basis of property similarity according to the ranges of HL_N , K_{OW} and MW, using a 2:1 ratio (i.e., 421 compounds in the training set, 211 compounds in the prediction set). The modelled data were transformed into logarithmic units ($\text{Log}_{10} HL_N$) in order to linearize the distribution of the response. The equation for this model can be expressed as:

$$\text{Log}HL_N = a_0 + a_1f_1 + a_2f_2 + \dots + a_n f_n + aMW + a\text{Log}Kow + e$$

where a_1, \dots, a_n are the regression coefficients calculated for the n descriptors, f_1, \dots, f_n are the molecular fragments, and e is the error term. This model was characterized by good fitting ($R^2 = 0.82$) and robustness ($Q^2 = 0.75$). The Mean Absolute Error calculated for the training set chemicals (MAE = 0.38 log units) was comparable with MAE calculated for the prediction set (0.45 log units) which suggests that the model is able to provide external predictions with similar accuracy as in the training set.

The applicability of this model is determined by the large set of heterogeneous fragments included in the equation; however, the dataset does not cover the domain of all the possible fragments included in organic chemicals. Moreover training and prediction sets included only a few substances with appreciable ionization at physiological pH and with MW > 600. Finally, as all the other models presented in this chapter, the BCFBAF model is not suitable for application on metals and organometallic compounds. These limitations in the applicability domain should be considered to avoid the generation of unreliable predictions.

3.1.2 IFS HL_N -QSAR

A second modelling study was conducted by Brown and colleagues (Brown et al. 2012) on the k_B (HL_N) dataset included in EPI Suite™. The new model was based on the Iterative Fragment Selection (IFS) method, which covered fragment generation, data splitting and model generation. The typology of fragments included in the model differs from those included in the EPI-QSAR model commented above, which had been pre-selected on the basis of a priori expert judgement (Arnot et al. 2009). In the IFS method, a pool of fragments is initially generated within the studied dataset by breaking all single and aromatic bonds with the exception of bonds with hydrogen. Through an automated iterative process of model fitting and cross-validation a final set of fragments (model descriptors) is selected (Brown et al. 2012). The IFS method was used to develop two multiple linear regression (MLR) models using two slightly different training sets depending on the splitting. The model named IFS-EPI was based on the same training set as the EPI-QSAR commented above, and the model IFS- HL_N was based on a new splitting generated with the IFS approach, keeping the 2:1 partition balance described in Sect. 3.1.1 (Brown et al. 2012).

The models proposed by Brown and colleagues have similar performances to the EPI-QSAR model; however, they are based on a lower number of fragments (i.e., 36–38 vs. 59) and they do not include $\log K_{OW}$ and MW as descriptors. The IFS-QSAR study included an analysis for the prediction of HL_{NS} for 25 IOCs contained in the studied dataset (i.e., acids with $pK_a < 9.5$ and bases with $pK_a > 5.5$). The MAE values calculated for the IOCs (0.50 and 0.40 for the training and the prediction set, respectively) were comparable to those calculated for the remaining neutral compounds (i.e., MAE training 0.39, MAE prediction 0.45). The fragments selected by the IFS method were similar to those included in the EPI-QSAR model. This convergence is important since it demonstrates that the same structural information selected a priori by expert judgement can be selected on an objective statistical basis from a larger pool of structural fragments.

3.1.3 QSARINS- HL_N QSAR

A third study was performed on the k_B (HL_N) dataset by Papa and colleagues (Papa et al. 2014). They applied a statistical approach to develop multiple HL_N -QSAR models, which were generated on the basis of different partitions of the k_B (HL_N) dataset into training and prediction sets (keeping the proportion 2:1). Differently from the EPI-QSAR (Arnot et al. 2009) and the IFS HL_N -QSAR (Brown et al. 2012) these new models were based on a limited number of theoretical molecular descriptors, which included global descriptors and fragments. The advantage of using global descriptors is that they encode for holistic structural information taking

Table 2 List of the QSAR models developed in Papa et al. (2014), summary of the number of compounds in training ($n^{\circ}\text{TR}$) and prediction sets ($n^{\circ}\text{PR}$), and of the main parameters calculated to quantify fitting (R^2) and internal/external validation (Q^2_{loo} , Q^2_{ext} , RMSE training (RMSE TR), RMSE prediction (RMSE PR))

Equation	$n^{\circ}\text{TR}$	$n^{\circ}\text{PR}$	R^2	Q^2_{loo}	Range	RMSE	RMSE
					$Q^2_{\text{ext}}^a$	TR	PR
(M1) $\text{Log HL}_N = -4.081 + 1.082 \text{VAdjMat} - 0.122 \text{gmax} - 0.205 \text{nHBacc} + 0.119 \text{nX} - 0.116 \text{SaaaC} + 0.387 \text{FP503} + 2.294 \text{FP29} - 0.666 \text{minHBd} + 0.241 \text{ndSCH}$	421	211	0.7	0.73	0.76–0.77	0.6	0.56
(M2) $\text{Log HL}_N = -3.883 + 0.400 \text{nX} + 1.052 \text{VAdjMat} - 0.111 \text{gmax} - 0.147 \text{nHBa} - 0.007 \text{ATSm4} + 0.149 \text{MDEC-11} - 0.853 \text{minHBd} - 0.095 \text{SaaaC} - 0.188 \text{nHCHnX}$	405	227	0.8	0.73	0.76–0.77	0.59	0.56
(M3) $\text{Log HL}_N = -4.19 + 1.058 \text{VAdjMat} - 0.254 \text{MAXDP} + 0.112 \text{nX} - 0.154 \text{nHBacc} + 0.154 \text{MDEC-11} + 0.464 \text{FP362} - 0.141 \text{naaaC} - 0.804 \text{minHBd} - 0.311 \text{FP376}$	421	211	0.8	0.74	0.75	0.57	0.58

^aBased on the calculation of the parameters Q^2_{F1} , Q^2_{F2} and Q^2_{F3} (Papa et al. 2014)

into account the whole molecular structure and the various intermolecular interactions (Todeschini and Consonni 2000), which cannot be captured by descriptors characterizing the presence and the chemical composition of isolated fragments.

The equations and performance of the best three models developed in this study are reported in Table 2. These models are also available in the QSARINS-Chem (Gramatica et al. 2014) module in the software QSARINS (Gramatica et al. 2013).

All the models have good statistical performance and adequate ability to fit the training sets and predict external data points ($Q^2_{\text{ext}} > 0.75$, and similar values of Residual Mean Squared Errors (RMSE) calculated for each training and the related prediction set). Statistical performances reported in Table 2 are comparable to those reported for the EPI- and the IFS-QSARs. However, the main advantage of the new HL_N -QSARs (i.e. M1, M2 and M3 in Table 2) is the lower complexity since they are based on 9 molecular descriptors, which is about 1/5 of the number of fragments used to develop the EPI-QSAR (Arnot et al. 2009).

Three molecular descriptors recurred in all three models and were the most relevant for modelling the selected response. These were the vertex adjacency index (VAdjMat), the number of halogen atoms (nX) and the minimum electrotopological-state energy for the hydrogen bond donor (minHBd). The first two descriptors have a positive correlation with the response and encode for

information about molecular dimension, and presence of halogen atoms, which are related to hydrophobicity and persistence. The latter, is inversely correlated to the response and describes the ability of the chemical to participate in intramolecular interactions, such as hydrogen bonds responsible of hydrophilicity. Other variables, which encode for information about the electrotopological state, were alternatively selected in the models. These are, respectively, the maximum electrotopological state (g_{max}) (Kier and Hall 1999), the maximum electrotopological positive variation (MAXDP), the count of the atom type E-state ::C: (naaaC) and the sum of atom type E-state ::C: (SaaaC).

Descriptors selected in the models suggested that HL_N is associated to the ability of the chemicals to participate in non-covalent intramolecular interactions, i.e., limited reactivity of the chemicals with the surrounding environment is associated to longer HL_N . This is the case of chemicals characterized by large, hydrophobic, halogenated structures, with one or more aromatic rings. Large, non-aromatic ring systems are also included in this group of chemicals, such as hexabromocyclododecane and chlordane, as well as long aliphatic chains. On the contrary, the increasing presence of polar and ionisable groups, as well as the number and variety of reactive functional substitutes simultaneously present in the molecule, is associated with shorter HL_N (i.e., faster biotransformation rates).

Finally, predictions generated by M1–M3 QSARINS models and the EPI-QSAR were combined (i.e., averaged predictions) in order to improve the quality of single model predictions and possibly to enlarge the applicability domain than individual models (Zhu et al. 2008), and therefore the reliability of predictions. Results reported by Papa et al. (2014) demonstrated that the combinatorial approach improved the correct prediction of slow and very slow biotransformed compounds, thereby reducing the number of possible false negatives that may be predicted when using the individual models separately in screening procedures. The aforementioned QSARs (Sects. 3.1.1–3.1.3) were developed following OECD QSAR guidance (OECD 2007).

3.1.4 Abraham Solvation Descriptors HL^{diss} QSAR

Another QSAR was developed by Kuo and Di Toro (2013) and this addressed the prediction of biotransformation of neutral and weakly polar compounds by using some Abraham Solvation descriptors. Abraham descriptors (Todeschini and Consonni 2000) are calculated directly from the molecular structure and describe chemical partitioning as a linear combination of different intermolecular interactions. In this paper the descriptors E, S, A, B, V quantified respectively dispersive and polarization interactions, dipolar interactions, H-bond donation, H-bond acceptance and molar volume relating to the energy required for cavity formation.

In this QSAR the empirically-based k_B values were transformed to generate HLs based on estimates of the freely dissolved chemical concentrations in the fish. The

basic assumption is that chemicals more accessible for interactions with the enzymes are those freely dissolved, i.e., not bound to storage lipids or bulk protein phases. Thus the whole body biotransformation half-life for the dissolved fraction HL^{diss} can be expressed as:

$$HL^{\text{diss}} = HL\Phi_{\text{fish}}$$

Where Φ_{fish} is the freely dissolved fraction of chemical in the fish. HL^{diss} could be calculated for 424 chemicals because of the required information on fish composition necessary to calculate Φ_{fish} . The Abraham approach (i.e., based on the calculation of the Abraham descriptors) was then applied to model HL^{diss} values of 64 molecules randomly chosen out of the 424 chemicals with empirically derived HL^{diss} values. The equation of this model is reported as follows:

$$\text{Log}HL^{\text{diss}} = -0.6(\pm 0.3) + 2.2(\pm 0.3)B - 2.1(\pm 0.2)V$$

The model is based on the descriptors B and V , which encode respectively for hydrogen bond acceptance and molar volume relating to the energy required for cavity formation. Even though the performances of the HL^{diss} -QSAR are slightly inferior to other models presented in this chapter, this model is still satisfactory in terms of fitting and predictivity ($R^2 = 0.70$; RMSE training = 0.71; RMSE prediction = 0.71). Other than some limitations related to the domain of applicability of this model (i.e., the applicability domain was not quantified, the model is based on only two descriptors and has been trained on a smaller training set in comparison to other HL_{N} -QSARs), it represents a relatively simple approach to predict biotransformation half-lives on the basis of chemical bioavailability in the fish.

3.2 Models for Predicting Biotransformation Rates of Organic Chemicals in Humans

Following the methods developed for estimating biotransformation rate constants for fish, Arnot and colleagues (Arnot et al. 2014) developed an approach to derive whole body in vivo biotransformation half-life (HL_{B}) estimates from measured total elimination half-lives (HL_{T}) in human adults (e.g., Obach et al. 2008). The HL_{T} dataset is composed of 1105 heterogeneous organic compounds of measured and estimated adult total elimination half-lives, primarily for pharmaceuticals (80% of the dataset) and well known environmental contaminants (20% of the dataset). All of these data were collected from peer-reviewed sources that had been reported with some data quality assurance methods. Four HL_{B} datasets were derived from the empirically-based HL_{T} values and different parameterizations (assumptions) of a

1-compartment mass balance model for humans. Uncertainty analysis was also included in the HL_B estimates in this study (Arnot et al. 2014). Half-lives were transformed to base 10 logarithmic prior to QSAR modelling.

3.2.1 IFS $HL_{T, B}$ -QSAR

Five externally validated QSARs were developed for five half-life datasets (splits performed at 50:50 proportion) using the IFS QSAR approach (Brown et al. 2012; Arnot et al. 2014) and multiple linear regression. Satisfactory values of $R^2 > 0.70$ were reported for all the models with ranges of RMSE from 0.45 to 0.49 in the training sets and from 0.70 to 0.75 in the prediction sets. The HL_T QSAR and the best among the HL_B QSARs were developed using 63 and 62 molecular fragments respectively. In general, those fragments encoding for halogenation, presence of carbon-carbon double bonds, aromatic and aliphatic carbons and non-aromatic nitrogen were associated to increasing HL s.

3.2.2 QSARINS- $HL_{T, B}$ QSAR

Recent works of Papa and colleagues (Papa et al. 2016) were also focused on the development of QSAR models for the prediction of HL_T and HL_B on the basis of the five datasets created by Arnot et al. (2014) described above. Multiple Linear Regression (MLR) models were generated by the Ordinary Least Squares (OLS) method and variable selection by Genetic Algorithm in QSARINS (Gramatica et al. 2013).

Performances calculated for the five models are summarized in Table 3.

Particular attention was paid to the quality of the models by verifying robustness, external predictivity, and applicability domain, while looking for the best interpretability of the descriptors. It is interesting to note that RMSE values calculated for the training and the prediction sets are well balanced. Moreover, the external

Table 3 List of the QSAR models developed in Papa et al. (2016), number of compounds in training ($n^{\circ}TR$) and prediction sets ($n^{\circ}PR$) and summary of the main parameters calculated to quantify fitting (R^2) and internal/external validation (Q^2_{100} , Q^2_{ext} , RMSE training, RMSE prediction)

Mod.	$n^{\circ}TR$	$n^{\circ}PR$	R^2	Q^2_{100}	Range Q^2_{ext}	RMSE training	RMSE prediction
HLT	552	553	0.78	0.77	0.74-0.75	0.62	0.66
HLB1	505	506	0.77	0.76	0.75	0.64	0.67
HLB2	507	508	0.79	0.77	0.76	0.63	0.66
HLB3	467	468	0.79	0.78	0.75-0.76	0.62	0.68
HLB4	470	470	0.8	0.79	0.76	0.62	0.69

predictivity of these models, which have been developed using 9 molecular descriptors each, is better (RMSE Prediction range from 0.66 to 0.69) than QSARs developed by the ISF method (Arnot et al. 2014) based on more than 60 descriptors (RMSE Prediction range from 0.70 to 0.75).

The most relevant descriptors for *HL*-QSARs, reported in Table 3, are the sum of atom-type E-State:-Cl (SsCl), the average Broto-Moreau autocorrelation of lag 7 or 8, weighted by polarizabilities (AATS7p or AATS8p) and the number of Halogen atoms (nX). It is interesting to note that these recurrent variables are similar to those selected in the fish models described above (Papa et al. 2014). According to these findings, biotransformation potential seems to be influenced, and reduced, mainly by the presence of halogen atoms covalently bonded to carbon atoms, as well as by the presence of polar atoms on large molecules (e.g., polybrominated diphenyl ethers, polychlorinated biphenyls, polychlorinated dibenzodioxins and polychlorinated dibenzofurans).

These observations are in line with other studies where covalent bonds between aromatic carbon and halogen atoms, in particular chlorine, are described as very stable and possibly enhancing persistence and biopersistence of chemicals (Meylan et al. 2007; Howard and Muir 2011).

The utility of these QSARs was demonstrated by predicting biotransformation half-lives in humans and fish (Papa et al. 2016) for over 1300 Pharmaceuticals and Personal Care Products (PPCPs). The information obtained from the biotransformation models was used to refine the screening of their intrinsic potential behaviour as Persistent, Bioaccumulative, and Toxic compounds (i.e., PBTs) performed in former studies (Cassani and Gramatica 2015; Sangion and Gramatica 2016). The PBT screening was obtained by applying two different QSAR models (US EPA 2006; Papa and Gramatica 2010), and led to the creation of two priority lists for personal care products and pharmaceuticals (Cassani and Gramatica 2015; Sangion and Gramatica 2016).

Papa and colleagues (Papa et al. 2016) performed Principal Component Analysis (PCA) to combine fish and human *HLs* predicted by different models, and to project the studied PPCPs in a new multidimensional space (Fig. 2a and b).

The direction of the loadings (i.e., the weights of the original variables in the principal components, which are represented as segments with origin in 0 in Fig. 2b) indicates that the compounds are ranked from left to right according to their increasing *HLs* (i.e., increasing biopersistence).

PC1 distinguishes between PPCPs which have fast (small squares on the left side of Fig. 2a) or slow (rounds on the right side of Fig. 2a) biotransformation. PC2 separates the original variables (loadings) in two main groups (Fig. 2b) depending on the organism (i.e., human and fish).

In addition, PPCPs are distinguished (i.e., whole and empty symbols) accordingly to results of the PBT screening performed by Cassani and Gramatica (2015) and Sangion and Gramatica (2016).

PPCPs, which were previously screened as PBTs and included in the former priority lists, are indicated in Fig. 2a as whole symbols, while empty symbols are PPCPs, which were screened as of no relevant concern for their PBT properties.

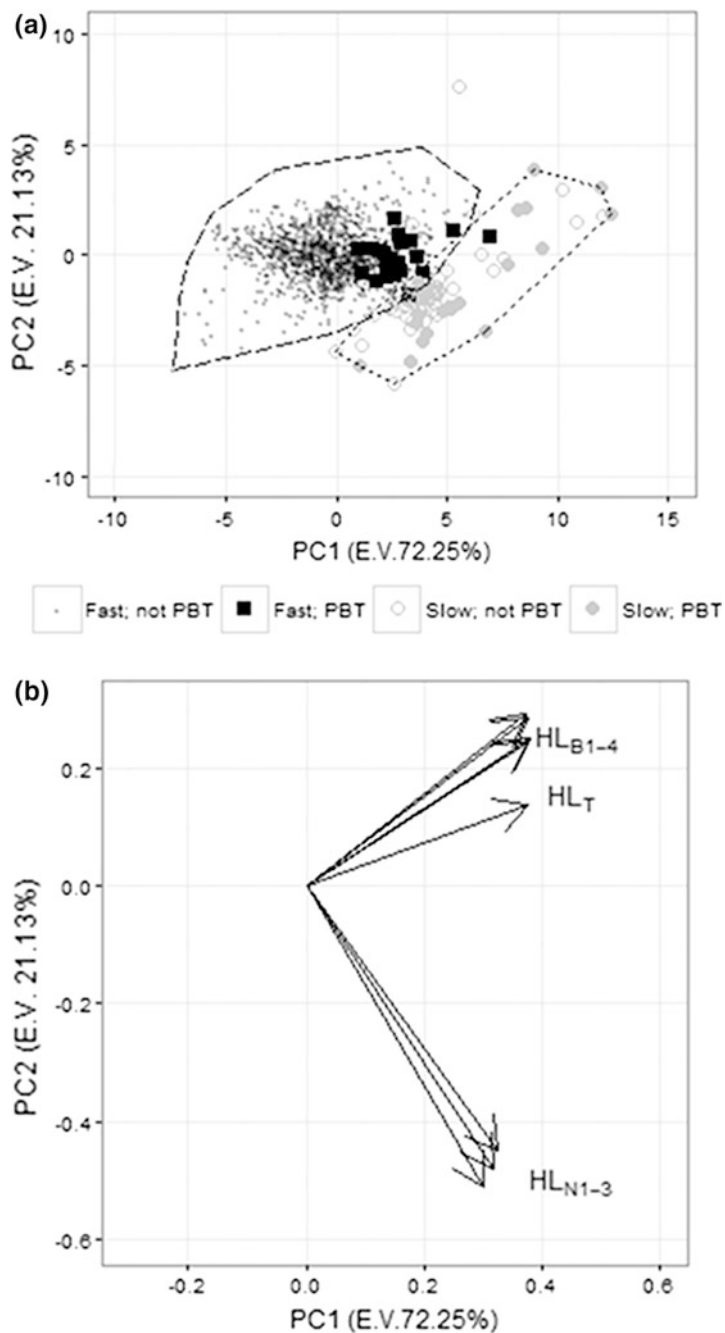


Fig. 2 a Plot of PC1 and PC2 scores from the PCA performed on fish and human HLs. The dotted line on the right side of the plot includes slowly biotransformed compounds with HLs > 10 days. The dotted line on the left side includes easily biotransformed compounds (i.e., HL < 10 days). b Plot of PC1 and PC2 loadings from the PCA performed on fish and human HLs

Therefore, using this combination between the PCA results based on predicted biotransformation *HLs* and the PBT screening, Papa and colleagues (Papa et al. 2016) were able to refine the earlier results obtained by Cassani and Gramatica (2015) and Sangion and Gramatica (2016). In fact they could highlight easily biotransformed PPCPs, which were previously predicted as PBTs (possible overestimation of the PBT behaviour), and slowly biotransformed PPCPs, which were previously predicted as non-PBTs. In particular, some PPCPs from the latter group, which had no agreement in predictions calculated by the models used to perform the PBT screening, were highlighted by Papa et al. (2016) as new priority compounds on the basis of the refinement of the PBT assessment.

3.3 *Key Points and Limitations in the Development and Application of QSAR Models*

Scientists and regulators have recently summarized the key points for the correct development of QSARs to increase transparency, harmonize calculations, provide better usability of the results and increase confidence in predictions generated by these tools (Gramatica 2007; OECD 2007; Fourches et al. 2010; Gramatica et al. 2012). The availability and correct use of (i) chemical structures, (ii) molecular descriptors which encode for the structural information, (iii) experimental data, which need to be numerically sufficient to generate statistically robust models, (iv) statistical procedures, are serious limiting factors which may strongly affect the final quality and application of the models.

The scarcity and the quality of experimental datasets, which are mostly collections of data scattered in the literature (Arnot et al. 2008b, 2014), and therefore are characterized by noise (uncertainty) due to experimental variability, are typical issues in the development of QSARs for biotransformation related processes. For instance, experimental variability was highlighted by Pirovano et al. (2015) as possible reason for the scarce performance of models derived for metabolic constants. Moreover, as specified in Sects. 3 and 3.2, the curation of empirically derived biotransformation data was the basic assumption of the work of Arnot and colleagues (Arnot et al. 2008b, 2014) in order to increase the quality of the modelled response and the accuracy of the derived QSAR models.

Another issue is the potential bias in the structural domain of biotransformation datasets representative for specific classes of compounds, such as pharmaceuticals in the human biotransformation dataset (Arnot et al. 2014) (Sect. 3.2), which may be not relevant for the structural features related to the biotransformation behaviour of other substances. Furthermore, the current lack of overlapping biotransformation datasets measured in different systems and conditions limits the possibility of extrapolation of QSAR results, i.e., from in vitro-to-in vivo or across different species.

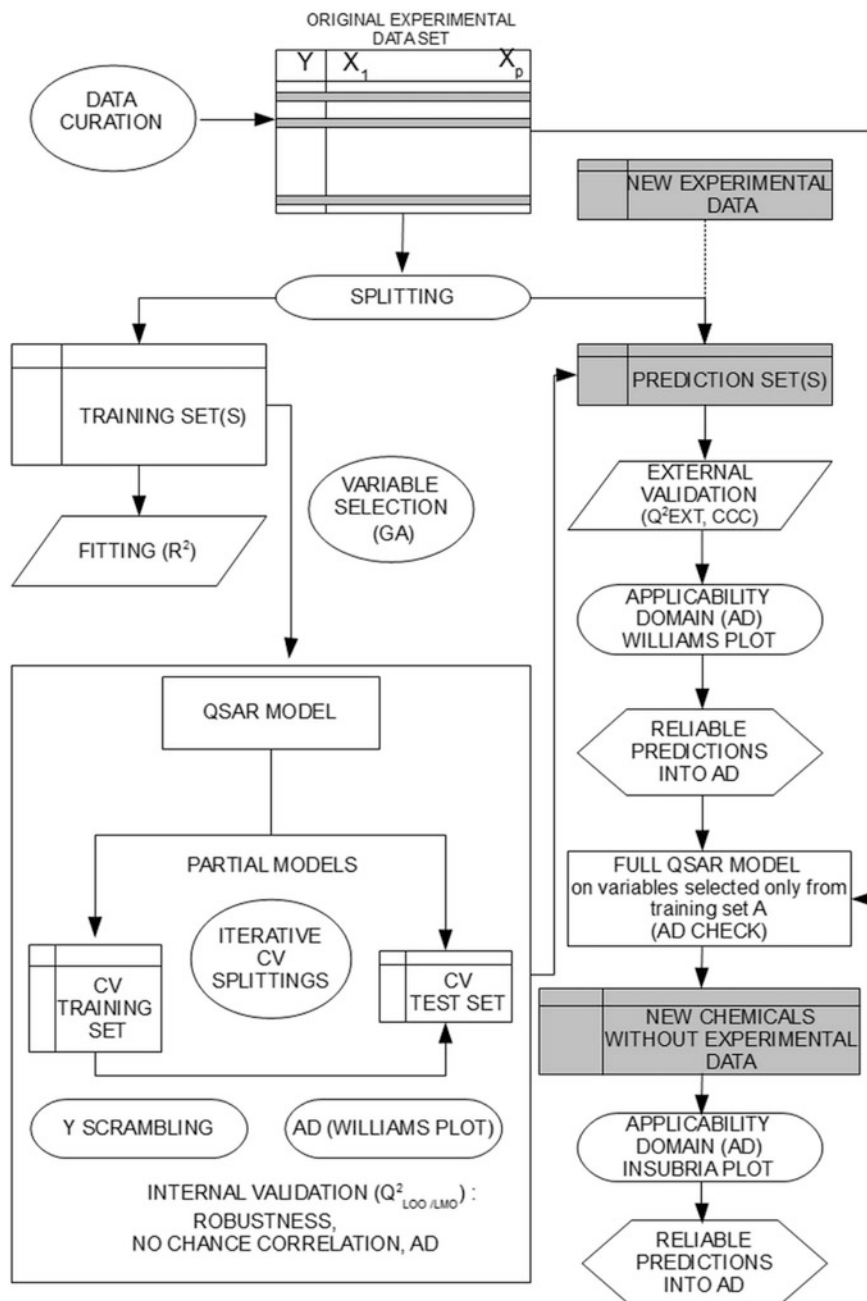


Fig. 3 Step-by-step procedure for the development of QSAR models. With permission from Gramatica et al. (2012)

These limitations impose a strict evaluation of the performance of the models, to guarantee the statistical reliability, and of the applicability domain intended as the structural space defined by the modelling descriptors, the range of the modelled response, and possibly the experimental/mechanistic context.

A step-by-step procedure for the development of QSAR models is reported in Fig. 3 (Gramatica et al. 2012). These steps are based the so-called “OECD principles for QSAR development and validation” (OECD 2007) and can be followed to generate statistically relevant and predictive models, independently of the modelled response (e.g., biotransformation) or the intended use (academic, commercial or regulatory).

4 In Vitro-to-In Vivo Calculations and In Silico Prediction

While *in vivo* experiments provide a more realistic representation of biotransformation in whole animals, costs and ethical concerns associated to animal testing represent a practical limitation for using *in vivo* methods for a large number of chemicals in a range of species. The desire to reduce costs and animal testing has provided the impetus for advancing *in vitro* biotransformation assays and IVIVE calculation methods, particularly for chemical hazard assessment. The pharmaceutical industry has been developing and refining methods for conducting *in vitro* biotransformation (metabolism) assays and using mathematical models for extrapolating *in vitro* rate estimates to *in vivo* clearance rates for several decades to aid in drug development screening (Wilkinson and Shand 1975; Rane et al. 1977; Wilkinson 1987; Brian Houston 1994; Iwatsubo et al. 1997; Obach et al. 1997; Austin et al. 2002; Riley et al. 2005; Fagerholm 2007; Rotroff et al. 2010; Obach 2011; Wetmore et al. 2014). The methods now used in environmental sciences (e.g., for fish Cravedi et al. 1999; Xing Han et al. 2007; Cowan-Ellsberry et al. 2008; Nichols et al. 2013) have evolved from the pharmaceutical sciences (e.g., for mammals). The biotransformation rate IVIVE methods are briefly described below.

The mathematical (*in silico*) models used to translate the *in vitro* measurements to *in vivo* tissue clearance are based on reaction rates, bioavailability (bound vs. “freely dissolved” unbound chemical), chemical partitioning, and blood/tissue flow rates. The extrapolation methods can include various measurement and QSAR models for the unbound (bioavailable) fraction (Zhu et al. 2013; Basant et al. 2016) and simplifying assumptions for the complex biological and chemical processes that occur during the biotransformation process, e.g., “well-stirred” or “parallel tube” models (Rane et al. 1977; Ito and Houston 2005; Fagerholm 2007; Yang et al. 2007). Although biotransformation can occur in multiple tissues and organs (example, gastrointestinal tract, gill, lung, skin, kidneys, blood, etc.), the liver is often assumed to be the organ primarily responsible for biotransformation in vertebrates. Therefore, the *in vitro* quantification of biotransformation is generally based on test systems which reflect the hepatic biotransformation processes, such as primary cell cultures (i.e., cells isolated from the liver), and subcellular fractions

(i.e., supernatant isolated from subcellular fractions such as cytosolic, microsomal, and S9 fractions after centrifugation at different speeds) (Weisbrod et al. 2009). In vitro rates can be determined from the rate of product (metabolite) formation and Michaelis-Menten kinetic parameters (i.e., V_{MAX}/K_M) and from rates of substrate depletion (i.e., rate of parent chemical loss over time) (Nichols et al. 2006, 2007). The in vitro intrinsic clearance rates ($CL_{INT, IN VITRO}$) are converted to in vivo intrinsic clearance rates ($CL_{INT, IN VIVO}$; $mL h^{-1} kg^{-1}$) using scaling factors. $CL_{INT, IN VIVO}$ is further extrapolated to the tissue or organ clearance from which the in vitro materials were obtained, (e.g., liver) using a mathematical model. For example, hepatic clearance (CL_H ; $mL h^{-1} kg BW^{-1}$) can be calculated using the well-stirred liver model as (Wilkinson and Shand 1975):

$$CL_H = \frac{Q_H f_U CL_{INT, INVIVO}}{(Q_H + f_U CL_{INT, INVIVO})}$$

where Q_H ($mL h^{-1} kg^{-1}$) is the rate of blood flow to the liver per kilogram of body weight, f_U (unitless) is the fraction of unbound chemical in the blood plasma ($f_{U,P}$) divided by the fraction of unbound blood in the incubation medium ($f_{U, IN VITRO}$). The CL_H estimates and other possible in vitro based estimates for other compartments (e.g., kidney) can be used to parameterize physiologically-based pharmacokinetic (PBPK) models to calculate kinetic and bioaccumulation parameters in various species (Nichols et al. 2007).

Assuming there is no extra-hepatic biotransformation, k_B can be calculated from CL_H and the steady state volume of distribution (VD_{SS} ; $mL kg^{-1}$) as:

$$k_B = \frac{CL_H}{VD_{SS}}$$

If there is extra-hepatic biotransformation occurring, then the aforementioned estimate is expected to underestimate the actual rate of whole body biotransformation. The k_B values can be used to parameterize 1-compartment bioaccumulation models. To-date, 1-compartment models have been more commonly applied for bioaccumulation hazard assessments (i.e., for fish), whereas PBPK models have been more commonly used in the pharmaceutical and veterinary industries and for human health assessments.

One limitation with in vitro assays for testing environmental contaminants is the relatively short lifespan of the enzymatic material in the test system (hours to a few days) and it can be difficult to accurately determine slower rates of biotransformation in these short-lived systems (Hutzler et al. 2015). More recently, the development of three-dimensional liver spheroids have shown promise, particularly for more slowly biotransformed chemicals because of the longer viability of the enzymes (Miranda et al. 2009; Wilk-Zasadna et al. 2015; Pinheiro et al. 2016). Another limitation of the in vitro assays for testing environmental contaminants is that most chemicals that have high bioaccumulation potential have high K_{OW} (i.e., very low water solubility) and it is technically challenging to dose and quantify the

required low chemical concentrations in the systems. A related confounding issue when performing *in vitro* analysis is the reduction of the nominal concentration due to loss from plates and vials caused by evaporation and/or adsorption phenomena (Kramer et al. 2010; Armitage et al. 2014). For this reason, the use of nominal concentrations to quantify the results of *in vitro* test may result in an overestimation of biotransformation activity (Sijm et al. 2007). Standardized methods for hepatocyte (Fay et al. 2014, 2015) and S9 (Johanning et al. 2012) assays have recently been developed for fish to aid in chemical hazard (bioaccumulation) and risk evaluations. Standardized methods for mammalian test systems would be valuable contributions for the regulatory process.

There are *in vitro* stability and biotransformation rate data for thousands of chemicals tested in mammalian models, (e.g., <https://www.ebi.ac.uk/chembl/>); however, because most of these data were, or are proprietary, many of the pertinent test details, i.e., test concentrations, are not available and the applicability of the data are uncertain. Almost all of these data are for pharmaceuticals or pharmaceutical candidates; there are relatively fewer *in vitro* biotransformation rate data for environmental contaminants. Furthermore, there are comparatively much fewer *in vitro* biotransformation rate data for fish (about 100 chemicals). However, as the regulatory need to evaluate chemicals continues to grow, it is expected that more *in vitro* biotransformation rate data for pharmaceuticals and commercial chemicals in ecological receptors will be generated.

Extrapolated estimates of *in vivo* liver clearance data (CL_H) for mammals have been used to develop and test *in silico* QSAR models for predicting CL_H from chemical structure (Hsiao et al. 2013; Li et al. 2009; Lombardo et al. 2014; Paixão et al. 2010; Pirovano et al. 2016; Schneider et al. 1999). One structural property that commonly remains relevant in these models is chemical hydrophobicity. There are no QSARs for CL_H in fish, likely because of the relatively fewer measured data currently available. Finally, it is worth mentioning that there are several models that predict whole body, total clearance in mammals (i.e., humans), e.g., Obach et al. (1997), Wajima et al. (2002), Jolivet and Ward (2005), Yap et al. (2006), McGinnity et al. (2007), Lavé et al. (2009), Yu (2010), Obach (2011), Demir-Kavuk et al. (2011), Berellini et al. (2012), Tonnelier et al. (2012), Gombar and Hall (2013), Arnot et al. (2014), Lombardo et al. (2014), Huang et al. (2015), Varma et al. (2015).

5 Conclusions

The estimation of *in vivo* biotransformation rates using *in silico* approaches is a complex process. The present chapter provides an overview on the approaches, which have been recently proposed to generate biotransformation-QSARs and to perform *in vitro*-*in vivo* extrapolations. The proposed examples show that some limitations to the application of *in silico* approaches still exist primarily related to the number and quality of the experimental data available to describe *in vitro* and

in vivo processes. However, some reliable QSARs are available for the prediction of whole body biotransformation of heterogeneous chemicals in fish and humans to screen large amount of chemicals and possibly to support and refine hazard-assessment procedures. This is particularly useful to reduce the possibility of overestimation of the bioaccumulation potential of chemicals that exhibit a high rate of biotransformation (i.e., short *HLs*).

The benefits eventually gained through the application of the in silico approaches summarised in this chapter (i.e., reduction of experimental costs and ethical concerns associated to animal testing, as well as possibility to include these predictions in green design of chemicals prior to synthesis), should call for further commitment of the scientific and regulatory community to improve the current methodologies of estimation, and to focus future experimental needs for the refinement of the models. The use of combinatorial and weighted approaches should also be actively pursued to maximise the information content in the available experimental data.

References

- Armitage, J. M., Wania, F., & Arnot, J. A. (2014). Application of mass balance models and the chemical activity concept to facilitate the use of in vitro toxicity data for risk assessment. *Environmental Science and Technology*, *48*, 9770–9779. doi:[10.1021/es501955g](https://doi.org/10.1021/es501955g).
- Arnot, J. A., Brown, T. N., & Wania, F. (2014). Estimating screening-level organic chemical half-lives in humans. *Environmental Science and Technology*, *48*, 723–730. doi:[10.1021/es4029414](https://doi.org/10.1021/es4029414).
- Arnot, J. A., & Gobas, F. A. P. C. (2003). A generic QSAR for assessing the bioaccumulation potential of organic chemicals in aquatic food webs. *QSAR & Combinatorial Science*, *22*, 337–345. doi:[10.1002/qsar.200390023](https://doi.org/10.1002/qsar.200390023).
- Arnot, J. A., & Gobas, F. A. P. C. (2006). A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms. *Environmental Reviews*, *14*, 257–297. doi:[10.1139/a06-005](https://doi.org/10.1139/a06-005).
- Arnot, J. A., Mackay, D., & Bonnell, M. (2008a) Estimating metabolic biotransformation rates in fish from laboratory data. *Environmental Toxicology and Chemistry*, *27*, 341–351. doi:[10.1897/07-310r.1](https://doi.org/10.1897/07-310r.1).
- Arnot, J. A., Mackay, D., Parkerton, T. E., & Bonnell, M. (2008b). A database of fish biotransformation rates for organic chemicals. *Environmental Toxicology and Chemistry*, *27*, 2263–2270. doi:[10.1897/08-058.1](https://doi.org/10.1897/08-058.1).
- Arnot, J. A., Meylan, W., Tunkel, J., Howard, P. H., Mackay, D., Bonnell, M., et al. (2009). A quantitative structure-activity relationship for predicting metabolic biotransformation rates for organic chemicals in fish. *Environmental Toxicology and Chemistry*, *28*, 1168. doi:[10.1897/08-289.1](https://doi.org/10.1897/08-289.1).
- Austin, R. P., Barton, P., Cockroft, S. L., Wenlock, M. C., Riley, R. J., & Al, A. E. T. (2002). The influence of nonspecific microsomal binding on apparent intrinsic clearance, and its prediction from physicochemical properties. *Drug Metabolism and Disposition*, *30*, 1497–1503.
- Barber, M. C. (2008). Dietary uptake models used for modeling the bioaccumulation of organic contaminants in fish. *Environmental Toxicology and Chemistry*, *27*, 755–777. doi:[10.1897/07-462.1](https://doi.org/10.1897/07-462.1).

- Basant, N., Gupta, S., & Singh, K. P. (2016). Predicting binding affinities of diverse pharmaceutical chemicals to human serum plasma proteins using QSPR modelling approaches. *SAR and QSAR in Environmental Research*, 27, 67–85. doi:10.1080/1062936X.2015.1133700.
- Berellini, G., Waters, N. J., & Lombardo, F. (2012). In silico prediction of total human plasma clearance. *Journal of Chemical Information and Modeling*, 52, 2069–2078. doi:10.1021/ci300155y.
- Borodina, Y., Sady, A., Filimonov, D., Blinova, V., Dmitriev, A., & Poroikov, V. (2003). Predicting biotransformation potential from molecular structure. *Journal of Chemical Information and Computer Sciences*, 43, 1636–1646. doi:10.1021/ci034078l.
- Brian Houston, J. (1994). Utility of in vitro drug metabolism data in predicting in vivo metabolic clearance. *Biochemical Pharmacology*, 47, 1469–1479. doi:10.1016/0006-2952(94)90520-7.
- Brown, T. N., Arnot, J. A., & Wania, F. (2012). Iterative fragment selection: A group contribution approach to predicting fish biotransformation half-lives. *Environmental Science and Technology*, 46, 8253–8260. doi:10.1021/es301182a.
- Burkhard, L. P. (2003). Factors influencing the design of bioaccumulation factor and biota-sediment accumulation factor field studies. *Environmental Toxicology and Chemistry*, 22, 351–360. doi:10.1002/etc.5620220216.
- Burkhard, L. P., Arnot, J. A., Embry, M. R., Farley, K. J., Hoke, R. A., Kitano, M., et al. (2012). Comparing laboratory and field measured bioaccumulation endpoints. *Integrated Environmental Assessment and Management*, 8, 17–31. doi:10.1002/ieam.260.
- Cassani, S., & Gramatica, P. (2015). Identification of potential PBT behavior of personal care products by structural approaches. *Sustainable Chemistry and Pharmacy*, 1, 19–27. doi:10.1016/j.scp.2015.10.002.
- Cowan-Ellsberry, C. E., Dyer, S. D., Erhardt, S., Bernhard, M. J., Roe, A. L., Dowty, M. E., et al. (2008). Approach for extrapolating in vitro metabolism data to refine bioconcentration factor estimates. *Chemosphere*, 70, 1804–1817. doi:10.1016/j.chemosphere.2007.08.030.
- Cravedi, J. P., Lafuente, A., Baradat, M., Hillenweck, A., & Perdu-Durand, E. (1999). Biotransformation of pentachlorophenol, aniline and biphenyl in isolated rainbow trout (*Oncorhynchus mykiss*) hepatocytes: Comparison with in vivo metabolism. *Xenobiotica*, 29, 499–509. doi:10.1080/004982599238506.
- de Wolf, W., Seinen, W., & Hermens, J. L. M. (1993). Biotransformation and toxicokinetics of trichloroanilines in fish in relation to their hydrophobicity. *Archives of Environmental Contamination and Toxicology*, 25, 110–117. doi:10.1007/BF00230720.
- Demir-Kavuk, O., Bentzien, J., Muegge, I., & Knapp, E.-W. (2011). DemQSAR: Predicting human volume of distribution and clearance of drugs. *Journal of Computer-Aided Molecular Design*, 25, 1121–1133. doi:10.1007/s10822-011-9496-z.
- Dimitrov, S., Pavlov, T., Veith, G., & Mekenyan, O. (2011). Simulation of chemical metabolism for fate and hazard assessment. I. Approach for simulating metabolism. *SAR and QSAR in Environmental Research*, 22, 699–718. doi:10.1080/1062936X.2011.623323.
- European Chemicals Agency. (2008). Guidance on information requirements and chemical safety assessment: QSARs and grouping of chemicals. *Guidance for Implementing Reach R*, 6, 134.
- European Union. (2006). Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/4. *Official Journal of the European Communities*, 1–520.
- European Union. (2009). Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products. *Official Journal of the European Union L*, 342, 342–359.
- European Union. (2012). Regulation (EU) No 528/2012 of the European Parliament and of the Council of 22 May 2012 concerning the making available on the market and use of biocidal products. *Official Journal of the European Communities L*, 269, 1–15.
- Fagerholm, U. (2007). Prediction of human pharmacokinetics—evaluation of methods for prediction of hepatic metabolic clearance. *Journal of Pharmacy and Pharmacology*, 59, 803–828. doi:10.1211/jpp.59.6.0007.

- Fay, K. A., Mingoia, R. T., Goeritz, I., Nabb, D. L., Hoffman, A. D., Ferrell, B. D., et al. (2014). Intra- and interlaboratory reliability of a cryopreserved trout hepatocyte assay for the prediction of chemical bioaccumulation potential. *Environmental Science and Technology*, *48*, 8170–8178. doi:[10.1021/es500952a](https://doi.org/10.1021/es500952a).
- Fay, K. A., Nabb, D. L., Mingoia, R. T., Bischof, I., Nichols, J.W., Segner, H., et al. (2015). Determination of metabolic stability using cryopreserved hepatocytes from rainbow trout (*Oncorhynchus mykiss*). *Current Protocols in Toxicology*, *65*, 4.42.1–4.42.29. doi:[10.1002/0471140856.tx0442s65](https://doi.org/10.1002/0471140856.tx0442s65).
- Fourches, D., Muratov, E., & Tropsha, A. (2010). Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of Chemical Information and Modeling*, *50*, 1189–1204.
- Gombar, V. K., & Hall, S. D. (2013). Quantitative structure-activity relationship models of clinical pharmacokinetics: Clearance and volume of distribution. *Journal of Chemical Information and Modeling*, *53*, 948–957. doi:[10.1021/ci400001u](https://doi.org/10.1021/ci400001u).
- Gramatica, P. (2007). Principles of QSAR models validation: Internal and external. *QSAR & Combinatorial Science*, *26*, 694–701.
- Gramatica, P., Cassani, S., & Chirico, N. (2014). QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. *Journal of Computational Chemistry*, *35*, 1036–1044. doi:[10.1002/jcc.23576](https://doi.org/10.1002/jcc.23576).
- Gramatica, P., Cassani, S., Roy, P. P., Kovarich, S., Yap, C. W., & Papa, E. (2012). QSAR modeling is not “push a button and find a correlation”: A case study of toxicity of (Benzo-) triazoles on Algae. *Molecular Informatics*, *31*, 817–835. doi:[10.1002/minf.201200075](https://doi.org/10.1002/minf.201200075).
- Gramatica, P., Chirico, N., Papa, E., Cassani, S., & Kovarich, S. (2013). QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. *Journal of Computational Chemistry*, *34*, 2121–2132. doi:[10.1002/jcc.23361](https://doi.org/10.1002/jcc.23361).
- Han, X., Nabb, D. L., Mingoia, R. T., & Ching-Hui, Y. (2007). Determination of xenobiotic intrinsic clearance in freshly isolated hepatocytes from rainbow trout (*Oncorhynchus mykiss*) and rat and its application in bioaccumulation assessment. *Environmental Science and Technology*, *41*, 3269–3276. doi:[10.1021/ES0626279](https://doi.org/10.1021/ES0626279).
- Hendriks, A. J., van der Linde, A., Cornelissen, G., & Sijm, D. T. (2001). The power of size. 1. Rate constants and equilibrium ratios for accumulation of organic substances related to octanol-water partition ratio and species weight. *Environmental Toxicology and Chemistry*, *20*, 1399–1420. doi:[10.1002/etc.5620200703](https://doi.org/10.1002/etc.5620200703).
- Howard, P. H., & Muir, D. C. G. (2011). Identifying new persistent and bioaccumulative organics among chemicals in commerce II: Pharmaceuticals. *Environmental Science and Technology*, *45*, 6938–6946. doi:[10.1021/es201196x](https://doi.org/10.1021/es201196x).
- Hsiao, Y. W., Fagerholm, U., & Norinder, U. (2013). In silico categorization of in vivo intrinsic clearance using machine learning. *Molecular Pharmaceutics*, *10*, 1318–1321. doi:[10.1021/mp300484r](https://doi.org/10.1021/mp300484r).
- Huang, W., Geng, L., Deng, R., Lu, S., Ma, G., Yu, J., et al. (2015). Prediction of human clearance based on animal data and molecular properties. *Chemical Biology & Drug Design*, *86*, 990–997. doi:[10.1111/cbdd.12567](https://doi.org/10.1111/cbdd.12567).
- Hutzler, J. M., Ring, B. J., & Anderson, S. R. (2015). Low-turnover drug molecules: A current challenge for drug metabolism scientists. *Drug Metabolism and Disposition*, *43*, 1917–1928.
- Ito, K., & Houston, J. B. (2005). Prediction of human drug clearance from in vitro and preclinical data using physiologically based and empirical approaches. *Pharmaceutical Research*, *22*, 103–112. doi:[10.1007/s11095-004-9015-1](https://doi.org/10.1007/s11095-004-9015-1).
- Iwatsubo, T., Hirota, N., Ooie, T., Suzuki, H., Shimada, N., Chiba, K., et al. (1997). Prediction of in vivo drug metabolism in the human liver from in vitro metabolism data. *Pharmacology & Therapeutics*, *73*, 147–171.
- Johanning, K., Hancock, G., Escher, B., Adekola, A., Bernhard, M. J., Cowan-Ellsberry, C., et al. (2012). Assessment of metabolic stability using the rainbow trout (*Oncorhynchus mykiss*) liver S9 fraction. *Current Protocols in Toxicology*, *1*, 14.10.14.10.1–14.10.28. doi:[10.1002/0471140856.tx1410s53](https://doi.org/10.1002/0471140856.tx1410s53).

- Jolivette, L. J., & Ward, K. W. (2005). Extrapolation of human pharmacokinetic parameters from rat, dog, and monkey data: Molecular properties associated with extrapolative success or failure. *Journal of Pharmaceutical Sciences*, *94*, 1467–1483. doi:10.1002/jps.20373.
- Kier, L. B., & Hall, L. H. (1999). *Molecular structure description: The electrotopological state*. Academic Press.
- Kim, J., Gobas, F. A. P. C., Arnot, J. A., Powell, D. E., Seston, R. M., & Woodburn, K. B. (2016). Evaluating the roles of biotransformation, spatial concentration differences, organism home range, and field sampling design on trophic magnification factors. *Science of the Total Environment*, *551*–552, 438–451. doi:10.1016/j.scitotenv.2016.02.013.
- Kramer, N. I., Busser, F. J. M., Oosterwijk, M. T. T., Schirmer, K., Escher, B. I., & Hermens, J. L. M. (2010). Development of a partition-controlled dosing system for cell assays. *Chemical Research in Toxicology*, *23*, 1806–1814. doi:10.1021/tx1002595.
- Kuo, D. T. F., & Di Toro, D. M. (2013). Biotransformation model of neutral and weakly polar organic compounds in fish incorporating internal partitioning. *Environmental Toxicology and Chemistry*, *32*, 1873–1881. doi:10.1002/etc.2259.
- Laue, H., Gfeller, H., Jenner, K. J., Nichols, J. W., Kern, S., & Natsch, A. (2014). Predicting the bioconcentration of fragrance ingredients by rainbow trout using measured rates of in vitro intrinsic clearance. *Environmental Science and Technology*, *48*, 9486–9495. doi:10.1021/es500904h.
- Lavé, T., Chapman, K., Goldsmith, P., & Rowland, M. (2009). Human clearance prediction: Shifting the paradigm. *Expert Opinion in Drug Metabolism and Toxicology*, *5*, 1039–1048. doi:10.1517/17425250903099649.
- Lech, J. J., & Bend, J. R. (1980). Relationship between biotransformation and the toxicity and fate of xenobiotic chemicals in fish. *Environmental Health Perspectives*, *34*, 115–131.
- Li, H., Sun, J., Sui, X., Liu, J., Yan, Z., Liu, X., et al. (2009). First-principle, structure-based prediction of hepatic metabolic clearance values in human. *European Journal of Medicinal Chemistry*, *44*, 1600–1606. doi:10.1016/j.ejmech.2008.07.027.
- Lillicrap, A., Springer, T., & Tyler, C. R. (2016). A tiered assessment strategy for more effective evaluation of bioaccumulation of chemicals in fish. *Regulatory Toxicology and Pharmacology*, *75*, 20–26. doi:10.1016/j.yrtph.2015.12.012.
- Lombardo, F., Obach, R. S., Varma, M. V., Stringer, R., & Berellini, G. (2014). Clearance mechanism assignment and total clearance prediction in human based upon in silico models. *Journal of Medicinal Chemistry*, *57*, 4397–4405. doi:10.1021/jm500436v.
- Long, A., & Walker, J. D. (2003). Quantitative structure-activity relationships for predicting metabolism and modeling cytochrome P450 enzyme activities. *Environmental Toxicology and Chemistry*, *22*, 1894–1899.
- Mackay, D. (1982). Correlation of bioconcentration factors. *Environmental Science and Technology*, *16*, 274–278. doi:10.1021/es00099a008.
- Mackay, D., Celsie, A. K. D., Arnot, J. A., & Powell, D. E. (2016). Processes influencing chemical biomagnification and trophic magnification factors in aquatic ecosystems: Implications for chemical hazard and risk assessment. *Chemosphere*, *154*, 99–108. doi:10.1016/j.chemosphere.2016.03.048.
- McGinnity, D. F., Collington, J., Austin, R. P., & Riley, R. J. (2007). Evaluation of human pharmacokinetics, therapeutic dose and exposure predictions using marketed oral drugs. *Current Drug Metabolism*, *8*, 463–479. doi:10.2174/138920007780866799.
- Mekenyan, O., Dimitrov, S., Pavlov, T., Dimitrova, G., Todorov, M., Petkov, P., et al. (2012). Simulation of chemical metabolism for fate and hazard assessment. V. Mammalian hazard assessment. *SAR and QSAR in Environmental Research*, *23*, 553–606. doi:10.1080/1062936X.2012.679689.
- Meylan, W., Boethling, R., Aronson, D., Howard, P., & Tunkel, J. (2007). Chemical structure-based predictive model for methanogenic anaerobic biodegradation potential. *Environmental Toxicology and Chemistry*, *26*, 1785–1792. doi:10.1897/06-579R.1.

- Miranda, J. P., Leite, S. B., Müller-Vieira, U., Rodrigues, A., Carrondo, M. J. T., & Alves, P. M. (2009). Towards an extended functional hepatocyte in vitro culture. *Tissue Engineering Part C*, *15*, 157–167. doi:[10.1089/ten.tec.2008.0352](https://doi.org/10.1089/ten.tec.2008.0352).
- Nichols, J. W., Fitzsimmons, P. N., & Burkhard, L. P. (2007). In vitro-in vivo extrapolation of quantitative hepatic biotransformation data for fish. II. Modeled effects on chemical bioaccumulation. *Environmental Toxicology and Chemistry*, *26*, 1304–1319. doi:[10.1897/06-259R.1](https://doi.org/10.1897/06-259R.1).
- Nichols, J. W., Huggett, D. B., Arnot, J. A., Fitzsimmons, P. N., & Cowan-Ellsberry, C. E. (2013). Toward improved models for predicting bioconcentration of well-metabolized compounds by rainbow trout using measured rates of in vitro intrinsic clearance. *Environmental Toxicology and Chemistry*, *32*, 1611–1622. doi:[10.1002/etc.2219](https://doi.org/10.1002/etc.2219).
- Nichols, J. W., Schultz, I. R., & Fitzsimmons, P. N. (2006). In vitro-in vivo extrapolation of quantitative hepatic biotransformation data for fish. I. A review of methods, and strategies for incorporating intrinsic clearance estimates into chemical kinetic models. *Aquatic Toxicology*, *78*, 74–90.
- Obach, R. S. (2011). Predicting clearance in humans from in vitro data. *Current Topics in Medicinal Chemistry*, *11*, 334–339. doi:[10.2174/156802611794480873](https://doi.org/10.2174/156802611794480873).
- Obach, R. S., Baxter, J. G., Liston, T. E., Silber, B. M., Jones, B. C., MacIntyre, F., et al. (1997). The prediction of human pharmacokinetic parameters from preclinical and in vitro metabolism data. *Journal of Pharmacology and Experimental Therapeutics*, *283*, 46–58.
- Obach, R. S., Lombardo, F., & Waters, N. J. (2008). Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds. *Drug Metabolism and Disposition*, *36*, 1385–1405. doi:[10.1124/dmd.108.020479](https://doi.org/10.1124/dmd.108.020479).
- OECD. (2012). Test No 305: Bioaccumulation in fish : Aqueous and dietary exposure. *Test No 305 Bioaccumulation Fish Aqueous Diet Expo Section*, *3*, 1–72. doi:[10.1787/2074577x](https://doi.org/10.1787/2074577x).
- OECD. (2007). Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models. *2*, 1–154. doi:[10.1787/9789264085442-en](https://doi.org/10.1787/9789264085442-en).
- Paixão, P., Gouveia, L. F., & Morais, J. A. G. (2010). Prediction of the in vitro intrinsic clearance determined in suspensions of human hepatocytes by using artificial neural networks. *European Journal of Pharmaceutical Sciences*, *39*, 310–321. doi:[10.1016/j.ejps.2009.12.007](https://doi.org/10.1016/j.ejps.2009.12.007).
- Papa, E., & Gramatica, P. (2010). QSPR as a support for the EU REACH regulation and rational design of environmentally safer chemicals: PBT identification from molecular structure. *Green Chemistry*, *12*, 836–843. doi:[10.1039/B923843C](https://doi.org/10.1039/B923843C).
- Papa, E., Sangion, A., Arnot, J. A., & Gramatica, P. (2016). Development of human biotransformation QSARs and application for PBT assessment refinement. *Food and Chemical Toxicology*. doi:[10.1016/j.fct.2017.04.016](https://doi.org/10.1016/j.fct.2017.04.016).
- Papa, E., van der Wal, L., Arnot, J. A., & Gramatica, P. (2014). Metabolic biotransformation half-lives in fish: QSAR modeling and consensus analysis. *Science of the Total Environment*, *470–471*, 1040–1046. doi:[10.1016/j.scitotenv.2013.10.068](https://doi.org/10.1016/j.scitotenv.2013.10.068).
- Peach, M. L., Liu, R., Pugliese, A., Wallqvist, A., & Nicklaus, M. C. (2012). Technology Review Computational tools and resources for metabolism-related property predictions. 1. Overview of publicly available (free and commercial) databases and software. *Future Medical Chemistry*, *4*, 1907–1932.
- Peters, R. H. (1983). *The ecological implications of body size*. Cambridge: Cambridge University Press.
- Pinheiro, P. F., Pereira, S. A., Harjivan, S. G., Martins, I. L., Marinho, A. T., Cipriano, M., et al. (2016). Hepatocyte spheroids as a competent in vitro system for drug biotransformation studies: Nevirapine as a bioactivation case study. *Archives of Toxicology*, 1–13. doi:[10.1007/s00204-016-1792-x](https://doi.org/10.1007/s00204-016-1792-x).
- Pirovano, A., Brandmaier, S., Huijbregts, M. A. J., Ragas, A. M. J., Veltman, K., & Hendriks, A. J. (2015). The utilisation of structural descriptors to predict metabolic constants of xenobiotics in mammals. *Environmental Toxicology and Pharmacology*, *39*, 247–258. doi:[10.1016/j.etap.2014.11.025](https://doi.org/10.1016/j.etap.2014.11.025).

- Pirovano, A., Brandmaier, S., Huijbregts, M. A. J., Ragas, A. M. J., Veltman, K., & Hendriks, A. J. (2016). QSARs for estimating intrinsic hepatic clearance of organic chemicals in humans. *Environmental Toxicology and Pharmacology*, 42, 190–197. doi:10.1016/j.etap.2016.01.017.
- Pirovano, A., Huijbregts, M. A. J., Ragas, A. M. J., & Hendriks, A. J. (2012). Compound lipophilicity as a descriptor to predict binding affinity (1/Km) in mammals. *Environmental Science and Technology*, 46, 5168–5174. doi:10.1021/es204506g.
- Rane, A., Wilkinson, G. R., & Shand, D. G. (1977). Prediction of hepatic extraction ratio from in vitro measurement of intrinsic clearance. *Journal of Pharmacology and Experimental Therapeutics*, 200, 420–424. doi:10.1016/0014-2999(77)90123-6.
- Riley, R. J., Mcginnity, D. F., & Austin, R. P. (2005). A unified model for predicting human hepatic, metabolic clearance from in vitro intrinsic clearance data in hepatocytes and microsomes. *Pharmacology*, 33, 1304–1311. doi:10.1124/dmd.105.004259.lenged.
- Rotroff, D. M., Wetmore, B. A., Dix, D. J., Ferguson, S. S., Clewell, H. J., Houck, K. A., et al. (2010). Incorporating human dosimetry and exposure into high-throughput in vitro toxicity screening. *Toxicological Sciences*, 117, 348–358. doi:10.1093/toxsci/ktq220.
- Sangion, A., & Gramatica, P. (2016). PBT assessment and prioritization of contaminants of emerging concern: Pharmaceuticals. *Environmental Research*, 147, 297–306. doi:10.1016/j.envres.2016.02.021.
- Schneider, G., Coassolo, P., & Lavé, T. (1999). Combining in vitro and in vivo pharmacokinetic data for prediction of hepatic drug clearance in humans by artificial neural networks and multivariate statistical techniques. *Journal of Medicinal Chemistry*, 42, 5072–5076. doi:10.1021/jm991030j.
- Segner, H. (2015). In vitro methodologies in ecotoxicological hazard assessment: The case of bioaccumulation testing for fish. *ATLA Alternatives to Laboratory Animals*, 43, P14–P16.
- Sevior, D. K., Pelkonen, O., & Ahokas, J. T. (2012). Hepatocytes: The powerhouse of biotransformation. *International Journal of Biochemistry & Cell Biology*, 44, 257–261. doi:10.1016/j.biocel.2011.11.011.
- Sijm, D. T. H. M., Rikken, M. G. J., Rorije, E., Traas, T. P., McLachlan, M. S., & Peijnenburg, W. J. G. M. (2007). Transport, accumulation and transformation processes. In C. J. van Leeuwen & T. G. Vermeire (Eds.), *Risk assessment of chemicals: An introduction, second* (pp. 73–158). Netherlands, Dordrecht: Springer.
- Todeschini, R., & Consonni, V. (2000). *Handbook of molecular descriptors*. Weinheim, Germany: Wiley-VCH Verlag GmbH.
- Tonnellier, A., Coecke, S., & Zaldívar, J. M. (2012). Screening of chemicals for human bioaccumulative potential with a physiologically based toxicokinetic model. *Archives of Toxicology*, 86, 393–403. doi:10.1007/s00204-011-0768-0.
- UNEP. (2016). In *Stockholm Convention 2008*. <http://chm.pops.int/TheConvention/Overview/TextoftheConvention/tabid/2232/Default.aspx>.
- US EPA. (2006). PBT Profiler. <http://www.pbtprofiler.net/>.
- Van der Linde, A., Jan Hendriks, A., & Sijm, D. T. H. M. (2001). Estimating biotransformation rate constants of organic chemicals from modeled and measured elimination rates. *Chemosphere*, 44, 423–435. doi:10.1016/S0045-6535(00)00213-7.
- van Leeuwen, C. J., & Vermeire, T. G. (Eds.). (2007). *Risk assessment of chemicals: An introduction*. Springer, Netherlands, Dordrecht: Second.
- Varma, M. V., Steyn, S. J., Allerton, C., & El-Kattan, A. F. (2015). Predicting clearance mechanism in drug discovery: Extended clearance classification system (ECCS). *Pharmaceutical Research*, 32, 3785–3802.
- Veith, G. D., DeFoe, D. L., & Bergstedt, B. V. (1979). Measuring and estimating the bioconcentration factor of chemicals in fish. *Journal of the Fisheries Research Board of Canada*, 36, 1040–1048. doi:10.1139/f79-146.
- Wajima, T., Fukumura, K., Yano, Y., & Oguma, T. (2002). Prediction of human clearance from animal data and molecular structural parameters using multivariate regression analysis. *Journal of Pharmaceutical Sciences*, 91, 2489–2499. doi:10.1002/jps.10242.

- Walker, C. H., Sibly, R. M., Hopkin, S. P., & Peakall, D. B. (2012). *Principles of ecotoxicology* (3rd ed.). Fourth Ed: CRC Press, Taylor and Francis Group.
- Weisbrod, A. V., Sahi, J., Segner, H., James, M. O., Nichols, J. W., Chultz, I. R. S., et al. (2009). The state of in vitro science for use in bioaccumulation assessments for fish. *Environmental Toxicology and Chemistry*, 28, 86–96. doi:[10.1897/08-015.1](https://doi.org/10.1897/08-015.1).
- Wetmore, B. A., Allen, B., Clewell, H. J., III, Parker, T., Wambaugh, J. F., Almond, L. M., et al. (2014). Incorporating population variability and susceptible subpopulations into dosimetry for high-throughput toxicity testing. *Toxicological Sciences*, 142, 210–224. doi:[10.1093/toxsci/kfu169](https://doi.org/10.1093/toxsci/kfu169).
- Wilkinson, G. R. (1987). Clearance approaches in pharmacology. *Pharmacological Reviews*, 39, 1–47.
- Wilkinson, G. R., & Shand, D. G. (1975). A physiological approach to hepatic drug clearance. *Clinical Pharmacology and Therapeutics*, 18, 377–390.
- Wilk-Zasadna, I., Bernasconi, C., Pelkonen, O., & Coecke, S. (2015). Biotransformation in vitro: An essential consideration in the quantitative in vitro-to-in vivo extrapolation (QIVIVE) of toxicity data. *Toxicology*, 332, 8–19. doi:[10.1016/j.tox.2014.10.006](https://doi.org/10.1016/j.tox.2014.10.006).
- Yang, J., Jamei, M., Yeo, K. R., Rostami-Hodjegan, A., & Tucker, G. T. (2007). Misuse of the well-stirred model of hepatic drug clearance. *Pharmacology*, 35, 501–502. doi:[10.1124/dmd.106.013359](https://doi.org/10.1124/dmd.106.013359).This.
- Yap, C. W., Li, Z. R., & Chen, Y. Z. (2006). Quantitative structure-pharmacokinetic relationships for drug clearance by using statistical learning methods. *Journal of Molecular Graphics and Modelling*, 24, 383–395. doi:[10.1016/j.jmgm.2005.10.004](https://doi.org/10.1016/j.jmgm.2005.10.004).
- Yu, M. J. (2010). Predicting total clearance in humans from chemical structure. *Journal of Chemical Information and Modeling*, 50, 1284–1295. doi:[10.1021/ci1000295](https://doi.org/10.1021/ci1000295).
- Zhu, H., Tropsha, A., Fourches, D., Varnek, A., Papa, E., Gramatical, P., et al. (2008). Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena piriformis*. *Journal of Chemical Information and Modeling*, 48, 766–784. doi:[10.1021/ci700443v](https://doi.org/10.1021/ci700443v).
- Zhu, X. W., Sedykh, A., Zhu, H., Liu, S. S., & Tropsha, A. (2013). The use of pseudo-equilibrium constant affords improved QSAR models of human plasma protein binding. *Pharmaceutical Research*, 30, 1790–1798. doi:[10.1007/s11095-013-1023-6](https://doi.org/10.1007/s11095-013-1023-6).

Development of Monte Carlo Approaches in Support of Environmental Research

Alla P. Toropova, Andrey A. Toropov, Emilio Benfenati, Robert Rallo, Danuta Leszczynska and Jerzy Leszczynski

Abstract CORAL software (<http://www.insilico.eu/coral>) was developed to assist the computational research based on quantitative structure—activity relationships (QSAR). It has been successfully applied in a number of research projects. CORAL is based on molecular features extracted from the simplified input-line entry system (SMILES) by means of additional molecular features extracted from hydrogen suppressed molecular graphs. Among vital applications of CORAL are those related to the evaluation of environmental effects of various chemical compounds. Few such examples are discussed in this chapter. The toxicity of a large group of chemicals towards fish, rat, and quail was examined as endpoint for the QSAR analysis. This study resulted in model improvement. Based on the obtained results,

A.P. Toropova (✉) · A.A. Toropov · E. Benfenati
IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19,
20156 Milan, Italy
e-mail: alla.toropova@marionegri.it

A.A. Toropov
e-mail: andrey.toropov@marionegri.it

E. Benfenati
e-mail: emilio.benfenati@marionegri.it

R. Rallo
Departament d' Enginyeria Informatica iMatematiques, Universitat Rovirai Virgili,
Av. Països Catalans, 26, 43007 Tarragona, Catalunya, Spain
e-mail: robert.rallo@urv.cat

D. Leszczynska
Department of Civil and Environmental Engineering,
Interdisciplinary Nanotoxicity Center, Jackson State University,
1325 Lynch St, Jackson, MS 39217-0510, USA
e-mail: danuta.leszczynska@jsums.edu

J. Leszczynski
Department of Chemistry and Biochemistry,
Interdisciplinary Nanotoxicity Center, Jackson State University,
1400 J. R. Lynch Street, 17910, Jackson, MS 39217, USA
e-mail: jerzy@icnanotox.org

the mechanistic interpretation and domain of applicability of the models for the above-mentioned endpoints was suggested.

Keywords QSAR • CORAL • Fish toxicity • Rat toxicity • Quail toxicity

1 Introduction

Environmental sciences cover a broad space of areas. It is useful to define some sectors of this space. The topics of major concerns related to environment that we will address include three major compartments: (i) water; (ii) soil; and (iii) air. Both industry and agriculture activities significantly impact all compartments.

The environmental impacts caused by industry and agriculture represent a complex problem and are also notably connected with the economical evolution and further societal advances. The obligation of modern natural sciences encompasses the protection of the biological diversity, and the improvement of the life conditions for our society. Though the last task seems to be the most important for the human beings, it can not be accomplished without proper addressing the environmental and biological related endpoints.

In order to survive and advance the modern society has to develop valuable and long-range strategies to tackle all of these challenges. The impacts of industry and agriculture should be explored and evaluated through large set outcomes. They could be transferred to a volume of scientific data that forms “endpoints”—variables that could be quantified and examined.

The focus of this book chapter is the discussion of such data—three endpoints related to water, soil, and air. This provides a guide for broader investigations, including the use of more experimental data when available. In this chapter, we will present predictive models for three endpoints reflecting environmental impacts upon: (i) fish (water), (ii) rats (soil), and (iii) birds (air).

The data is arranged as follows:

Water: The toxicity of 568 industrial organic compounds expressed as 50% lethal concentration (LC50) for Fathead minnow;

Soil: The lethal rat toxicity expressed as negative decimal logarithm of the lethal dose in mg/kg (pLD50) for 525 compounds described in the literature;

Air: Toxicity towards quail of 115 compounds, expressed as the decimal logarithm $\log(1/C)$, where C is the concentration, in mmol/kg, expressed as LC50-96 h, the dose that kills 50% of quail population in 96 h.

Every year the list of compounds, used by the industry or in agriculture, becomes longer. The experimental assessment of each new compound is impossible. Under such circumstances, one needs a reliable tool to estimate possible damage that each new compound could cause. It should be done without expensive experiments.

There are a number of computational techniques developed in the last 20 years that might provide estimation of the environmental impacts of chemical

compounds. The CORAL software represents a numerical implementation of such techniques. The main principle that was applied when this software was developed was to properly address so-called correlation weights for various molecular features extracted from the simplified molecular input-line entry systems (SMILES) (Weininger 1988, 1990; Weininger et al. 1989). The application of Monte Carlo method allows providing the numerical data for these correlation weights.

CORAL is the abbreviation derived from “CORrelations And Logic” terms. The software generates both the structural descriptors and quantitative structure—property/activity relationship (QSPRs/or QSARs) models. In the first approximation, the structural descriptors are calculated based on the SMILES structural format, which is a very compact and useful way to describe the chemical in a string of characters, related to the atoms and bonds. CORAL uses SMILES according to the following scheme:

1. Each SMILES is converted into a group of attributes. An attribute can represent a fragment of SMILES line. This kind of attribute is a local one. Furthermore, a descriptor related to some molecular features such as, the presence or absence of various kinds of atoms (nitrogen, oxygen, chlorine, etc.) or various kinds of chemical bonds (simple covalent bond, double bond, triple bond) can represent an attribute.
2. The numerical data on so-called correlation weights of the above-mentioned attributes are calculated by the Monte Carlo method. For the training set, the correlation weights must give the maximal correlation coefficient between the sum of correlation weights of SMILES and the endpoint of interest.

Having the numerical data on the correlation weights, one can build up predictive models (QSPR or QSAR) as one-variable correlation between “endpoint—optimal descriptor”, where the optimal descriptor for a given SMILES is the sum of correlation weights of the above-mentioned attributes.

Unquestionably, the developed model should be checked up with external validation set. Such model can be useful if and only when the predictive potential of the model is confirmed for external validation set. Thus, similarly to other approaches, the CORAL model is sensitive to the distribution of available data into the training and validation sets.

The scheme is based solely on the characteristics of the molecular structure of the investigated compounds. Consequently, the CORAL approach retains definitions of the domain of applicability and mechanistic interpretation caused by the presence (and absence) of various SMILES attributes. Hence, the applicability domain and mechanistic interpretations can vary with different distributions of the data into the training and validation sets. This can be interpreted as a disadvantage of the approach because it affects the reproducibility of the statistical results. However, it can also be considered as an advantage because it allows comparing different distributions of the data into the training and validation sets and consequently, one can select the most reliable distribution. Such a distribution is characterized by a good prevalence for all important molecular features (attributes). In

addition, it gives possibility to detect uninformative molecular features with small prevalence.

Detailed description of the CORAL software is available on the Internet (<http://www.insilico.eu/coral>).

This chapter demonstrates and discusses the application of the above-mentioned approach to get data necessary for risk assessment of various chemicals, which are potential contaminants of water, soil, and air. This separation of the various environmental compartments (water, soil, air) provides a convenient and simple way to demonstrate the use of the Monte Carlo method to solve ecological tasks related to the various cases of risk assessment.

2 Method

2.1 Data

Water: The environmental effects of various chemicals dissolved in water have been studied by many groups. Here, we selected a toxicity study of a large group (568 industrial organic compounds). Their acute toxicity data expressed as 50% lethal concentration (LC50) for the juvenile stage of the Fathead minnow has been taken from the literature (Russom et al. 1997). The LC50 concentration is expressed in mmol/L. The experimental data was used to develop a QSAR model of toxicity. In the QSAR analysis, the endpoint was expressed as the negative decimal logarithm of the LC50, i.e., $-\log LC50$ or $pLC50$. The SMILES for examined compounds were generated with the ACD/ChemSketch software. QSAR models have been developed in work (Toropova et al. 2012) using data taken from (Russom et al. 1997).

Soil: The various chemicals have a significant effect on the health and population of soil organisms. To model such phenomena many experimental studies were performed on rats. We adopted the data on lethal rat toxicity data in mg/kg from the study published in the literature (Toropova et al. 2015a). The predicted endpoint represents the negative logarithm of the lethal dose ($pLD50$). The group of random distributions of all 525 compounds into the training, calibration, and validation sets was studied as the basis for building up model for $pLD50$. The SMILES were generated with ACD/ChemSketch software.

Air: Birds represent a fundamental group of the animals. The investigation of the influence of chemicals on birds can provide important clues on air quality. We selected a recent study on 114 chemical compounds performed on the quails. Decimal logarithm of the concentration, in mmol/kg, expressed as LC50-96 h, which is the dose that kills 50% of quails in 96 h was used as the endpoint to develop a QSAR relationship (Toropov and Benfenati 2007).

The computational studies described in this chapter were performed using commonly adopted rules. The experimental data were used to develop a QSAR

model and then to test its quality. It was done by splitting the data set into two general groups. For all examined endpoints, the splits of experimental data into training and test sets were carried out according to the following principles: (i) the range of the endpoint values is approximately the same for each sub-set; (ii) the splits are random; and (iii) the splits are not identical.

2.2 Optimal Descriptors

The CORAL software (<http://www.insilico.eu/coral>) provides a tool to build up QSAR models utilizing the Monte Carlo method. The possible representations of the molecular structure for these models are: (i) simplified molecular input-line entry system (SMILES) (Weininger 1988, 1990; Weininger et al. 1989), and (ii) molecular graphs. The CORAL software can convert SMILES into three kinds of the molecular graphs: (i) hydrogen suppressed graph (HSG) (Toropov et al. 2011); (ii) hydrogen filled graph (HFG) (Toropov et al. 2011); and graph of atomic orbitals (GAO) (Toropov et al. 2011).

The generalized form of a CORAL model could be described by the following one-variable equation (Fig. 1):

$$\text{Activity} = C_0 + C_1 \times \text{DCW}(T^*, N^*) \quad (1)$$

where, the Activity is an endpoint; C_0 and C_1 are regression coefficients; $\text{DCW}(T^*, N^*)$ is the optimal descriptor, which is a mathematical function of molecular features extracted from SMILES and/or graph. The numerical value of the $\text{DCW}(T^*, N^*)$ is calculated with so-called correlation weights (CW) of the above-mentioned molecular features. The numerical data on the correlation weights calculated with the Monte Carlo optimization is described in a series of studies (Toropov and Toropova 2002a, b; Toropova et al. 2016). There are two possible versions of the Monte Carlo optimization: (i) the classic scheme “training-calibration-validation”; and (ii) the balance of correlations, according to the scheme “training-invisible training-calibration-validation”. Thus, there is a number of possible ways to building up QSAR models using the CORAL software. The example of organization of the CORAL optimal descriptors is shown in Table 1. It follows the development of the computational model. The general scheme of planning of such tasks is depicted in Fig. 2. The CORAL program is well designed and performs all necessary tasks. Figure 3 displays a screenshot that provides an example of the organization of the optimal descriptor displayed in the Table 1.

This procedure was used to investigate the effects of chemicals for three major environmental compartments. Optimal descriptors were used in this work to build up models for the above-mentioned three endpoints. The following equations provide the details:

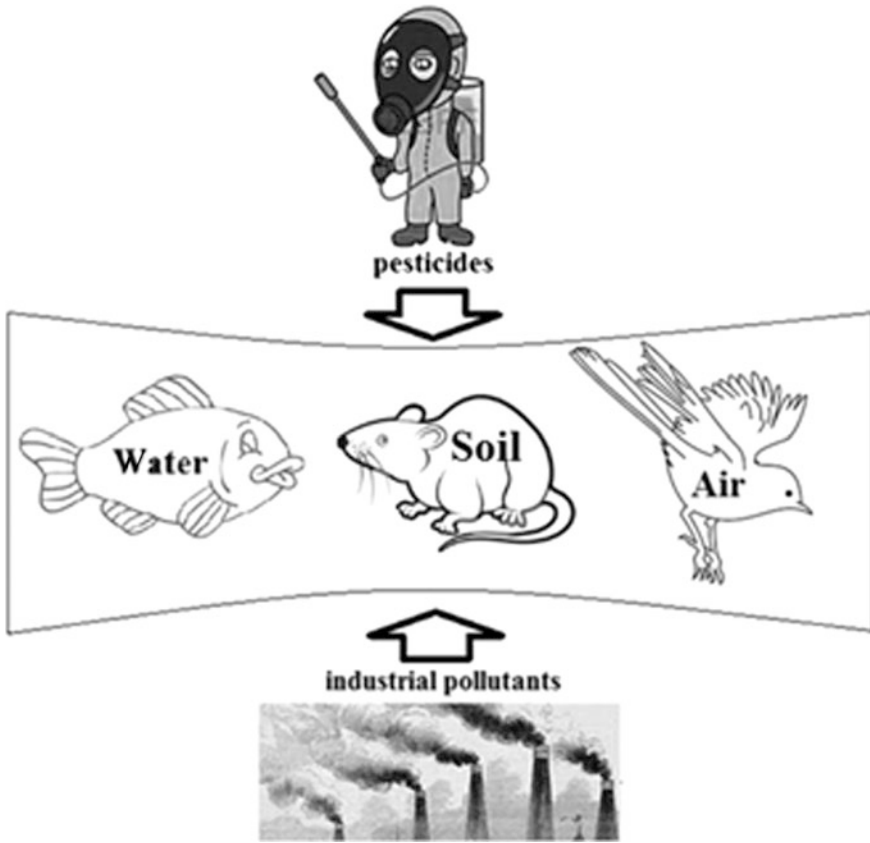


Fig. 1 General representation of main ecological problems

Water (lethal concentration $pLC50$ for Fathead minnow):

$$DCW(T, N) = \sum CW(NNC_k) + \sum CW(S_k) + \sum CW(SS_k) + \sum CW(SSS_k) \quad (2)$$

Soil (rat toxicity, lethal dose $pLD50$):

$$DCW(T, N) = \sum CW(NNC_k) + \sum CW(S_k) + \sum CW(SS_k) + \sum CW(SSS_k) + CW(BOND) + CW(NOSP) + CW(PAIR) \quad (3)$$

Air (quail toxicity, lethal dose $pLC50$):

$$DCW(T, N) = \sum CW(NNC_k) + \sum CW(S_k) + \sum CW(SS_k) + CW(BOND) + CW(NOSP) + CW(HALO) \quad (4)$$

Table 1 List of all available parameters for the CORAL software to build up QSPR/QSAR models by extracting molecular features from SMILES or molecular graph. Example of the selection of options for the case of the fish toxicity are indicated by grey background (selection of balance correlation)

SMILES	S_k	SS_k	SSS_k			Local attributes
	<i>BOND</i>	<i>NOSP</i>	<i>HALO</i>	<i>PAIR</i>		Global attributes
GS(HSG)	$EC0_k$	$EC1_k$	$EC2_k$	$EC3_k$		Local attributes
	$PT2_k$	$PT3_k$	$VS2_k$	$VS3_k$	NNC_k	
	<i>C3</i>	<i>C4</i>	<i>C5</i>	<i>C6</i>	<i>C7</i>	Global attributes
GS(HFG)	$EC0_k$	$EC1_k$	$EC2_k$	$EC3_k$		Local attributes
	$PT2_k$	$PT3_k$	$VS2_k$	$VS3_k$	NNC_k	
	<i>C3</i>	<i>C4</i>	<i>C5</i>	<i>C6</i>	<i>C7</i>	Global attributes
GS(GAO)	$EC0_k$	$EC1_k$	$EC2_k$	$EC3_k$		Local attributes
	$PT2_k$	$PT3_k$	$VS2_k$	$VS3_k$	NNC_k	

(*)There are two additional options: (i) classic scheme and (ii) balance of correlations; $EC0$, $EC1$, ..., $EC3$ are the extended connectivity of zero, first, ..., third orders (Toropov et al. 2013); $PT2$, $PT3$ are paths of length 2 and 3, respectively (Toropov et al. 2012); $VS2$, $VS3$ are valence shells of second and third orders, respectively (Toropov et al. 2012); NNC are nearest neighbours codes (Toropov et al. 2013); $C3$, $C4$, ..., $C7$ are global molecular attributes related to various rings (from three members till seven members rings); SS is SMILES system; GS is graph system; HSG , HFG , and GAO are hydrogen suppressed graph, hydrogen filled graph, and graph of atomic orbitals, respectively

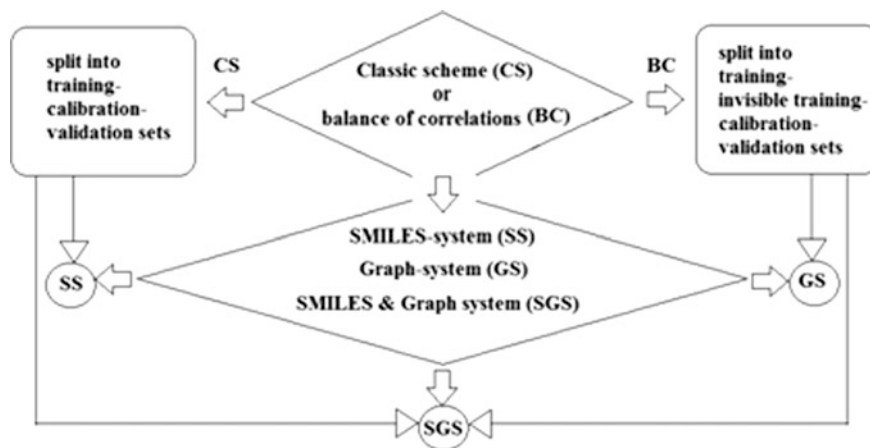


Fig. 2 The generalized scheme of planning of the study using CORAL software to build up a QSAR model

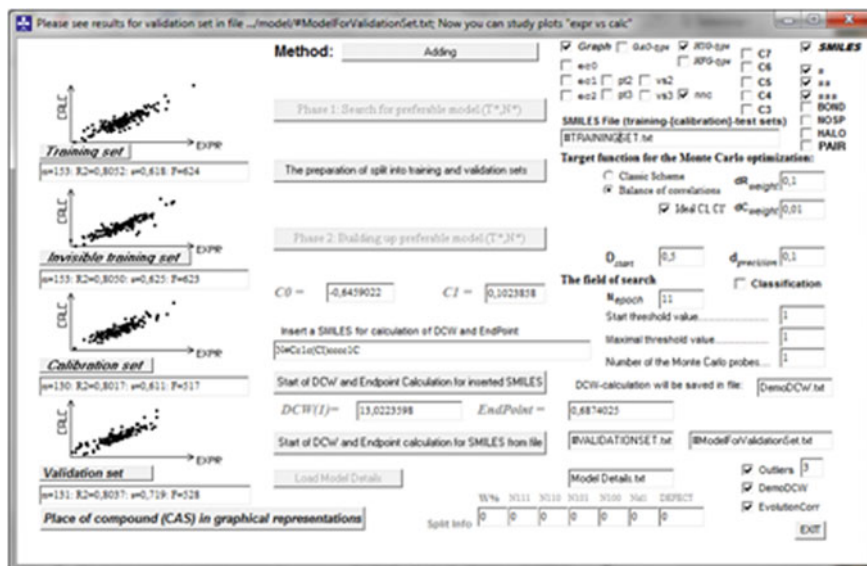
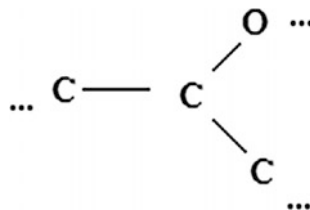


Fig. 3 Screenshot for method utilized to build up model for fish toxicity

Fig. 4 An example of k-th vertex



The nearest neighbouring codes NNC_k are described in the recent work (Toropov and Toropova 2002a):

$NNC[k] = 100 * N_{all} + 10 * N_{carbon} + N_{noncarbon}$ N_{all} , N_{carbon} , and $N_{noncarbon}$ are the total number of neighbors for kth vertex, the number of vertices which are carbon, and the number of vertices which are not carbon, respectively. For example, if kth vertex is as in Fig. 4, the $NNC[k] = 3 * 100 + 10 * 2 + 1 = 321$.

The S_k , SS_k , and SSS_k are local SMILES attributes described in work (Toropova et al. 2016);

For example SMILES = Clc1cccc1

$$S_k = (Cl, c, 1, c, c, c, c, 1);$$

$$SS_k = (Clc, c1, cc, cc, cc, cc, cc, c1);$$

$$SSS_k = (Clc1, c1c, ccc, ccc, ccc, ccc, cc1).$$

Table 2 Examples of definition for BOND, NOSP, HALO, and PAIR descriptors

Global attribute	Comment								
<i>BOND</i>	<p>The presence/absence of double ('='), triple ('#'), and stereo chemical ('@') bonds, e.g. if SMILES = "CVC = C\O"</p> <table border="1" style="display: inline-table; vertical-align: middle;"> <tr> <td>=</td> <td>#</td> <td>@</td> </tr> <tr> <td>1</td> <td>0</td> <td>0</td> </tr> </table> ⇒ BOND10000000	=	#	@	1	0	0		
=	#	@							
1	0	0							
<i>NOSP</i>	<p>Presence (absence) of nitrogen, oxygen, sulphur, and phosphorus, e.g. if SMILES = "CCC(O)CC"</p> <table border="1" style="display: inline-table; vertical-align: middle;"> <tr> <td>N</td> <td>O</td> <td>S</td> <td>P</td> </tr> <tr> <td>0</td> <td>1</td> <td>0</td> <td>0</td> </tr> </table> ⇒ NOSP01000000	N	O	S	P	0	1	0	0
N	O	S	P						
0	1	0	0						
<i>HALO</i>	<p>Presence (absence) of fluorine, chlorine, bromine, and iodine atoms, e.g. if SMILES = 'CICC(=O)CCI'</p> <table border="1" style="display: inline-table; vertical-align: middle;"> <tr> <td>F</td> <td>Cl</td> <td>Br</td> <td>I</td> </tr> <tr> <td>0</td> <td>1</td> <td>0</td> <td>0</td> </tr> </table> ⇒ HALO01000000	F	Cl	Br	I	0	1	0	0
F	Cl	Br	I						
0	1	0	0						
<i>PAIR</i>	<p>Simultaneous presence of two SMILES-components from the list: F, Cl, Br, I, N, O, S, P, #, =, and @; e.g. if SMILES = "CICC(=O)CCI" the following</p> <p>pairs will be extracted:</p> <table border="1" style="display: inline-table; vertical-align: middle;"> <tr> <td>+++Cl..O...</td> <td>i.e. 'Cl' and 'O'</td> </tr> <tr> <td>+++Cl..B2..</td> <td>i.e. 'Cl' and '='</td> </tr> <tr> <td>+++O...B2..</td> <td>i.e. 'O' and '='</td> </tr> </table>	+++Cl..O...	i.e. 'Cl' and 'O'	+++Cl..B2..	i.e. 'Cl' and '='	+++O...B2..	i.e. 'O' and '='		
+++Cl..O...	i.e. 'Cl' and 'O'								
+++Cl..B2..	i.e. 'Cl' and '='								
+++O...B2..	i.e. 'O' and '='								

BOND, NOSP, HALO, and PAIR are global SMILES attributes described in work (Toropova et al. 2016).

Table 2 contains examples of calculation for BOND, NOSP, HALO, and PAIR descriptors.

The $CW(NNC_k)$ are the correlation weights for nearest neighbouring codes in HSG (Toropov and Toropova 2002a). The $CW(S_k)$, $CW(SS_k)$, $CW(SSS_k)$, $CW(BOND)$, $CW(NOSP)$, $CW(HALO)$, and $CW(PAIR)$ are the correlation weights of the above SMILES attributes. T and N are parameters of the Monte Carlo optimization procedure (Toropov and Toropova 2002a, b; Toropova et al. 2016). T is the threshold used to classify structural attributes into two classes: (i) rare; and (ii) active. The rare attributes are blocked: their correlation weights are zero. The active attributes are involved in building up a model. N is the number of epochs of the optimization. If the $N \rightarrow \infty$, the overtraining becomes very plausible. $T = T^*$ and $N = N^*$ are the parameters that give best statistical quality of a model calculated with Eq. (1) for the calibration set (Toropova et al. 2016). T^* and N^* are

calculated by means of analysis of results of the Monte Carlo optimization in the ranges of $T \in (T_{\min}, T_{\max})$ and $N \in (1, N_{\max})$ Toropov et al. (2016, 2015b, 2013). In this work, the range of the threshold is (1, 3) and the range of the number of epochs is (1, 25).

There are certain conditions to be fulfilled when a QSAR approach is applied. QSAR is a statistically based model (Toropov et al. 2016, 2015b, 2013). This means that one should check up the developed equations related to a given endpoint using different splits of experimental data into the training and validation sets. The calculations with Eqs. (2)–(4) are carried out with the balance of correlations. Consequently, the training set examined in this work is organized as the training set (builder of a model), invisible training set (inspector of quality of a model during the Monte Carlo optimization), and calibration set (expert to define the moment of the rational stopping of the optimization process to avoid overtraining).

2.3 Tests of the CORAL Models and Definition of the Applicability Domain

It is important to evaluate the developed model for its predictability. This has been carefully done in the reported studies. Table 3 contains the list of criteria that estimate the reliability of the CORAL model from the probabilistic point of view. All these criteria have been checked in the three cases reported here.

2.4 Mechanistic Interpretation

The possibility to assess features related to the mechanisms of the studied phenomena represents a considerable advantage of the theoretical approaches we presented. The mechanistic interpretation of the developed QSAR models can be obtained from several runs of the Monte Carlo optimization. One can select three categories of attributes: (i) structural attributes, which have only positive values of the correlation weights. These are classified as promoters of the endpoint increase; (ii) structural attributes, which have only negative values of the correlation weights. These data are referred to as promoters of the endpoint decrease; and (iii) structural attributes, which have both positive and negative correlation weights in several runs of the Monte Carlo optimization. Obviously, these attributes should be considered as attributes with unclear role.

Table 3 The statistical criteria for probabilistic estimation of quality of (i) molecular feature extracted from SMILES or graph; (ii) split into the training and validation sets; (iii) domain of applicability

Criterion	Notes	Comment
$SA_{Defect} = \sum_{active} \sum P(SA) - P'(SA) $	Defect of structural attribute (SA), in terms of the inequality of probabilities in the training set $P(SA)$ and in calibration set $P'(SA)$. Defect = 0, if these probabilities are equal	This is estimation of significance of the structural attribute
$P(SA) = \frac{N_{set}(SA)}{N_{set}}$	$N_{set}(SA)$ is the number of SMILES which contain the given structural attribute; N_{set} is the total number of SMILES in a set (i.e. in training set or in calibration set)	
$SMILES_{defect} = \sum_{SA_{defect} \in SMILES} SA_{Defect}$	Defect of SMILES is sum of not blocked structural attributes	This estimation of reliability of prediction for the SMILES: if, defect is too large the prediction is problematic
$Split_{defect} = \sum_{SMILES \in Training} SMILES_{Defect}$		This criterion allows to select preferable split into the training and calibration sets
$SMILES_{defect} < 2 \times \overline{SMILES_{defect}}$		This criterion allows to select the domain of applicability

3 Results and Discussion

Each QSAR model is characterized by a set of statistical parameters. There is the need for broad tools to develop QSAR models in a facilitated way, which is flexible to different collections of data, and can be easily implemented. Possibly, the tools to be used should be freely available, for the broader dissemination, and they should wrap the calculation of the descriptors and the model algorithm into the same architecture.

In the past years we developed and updated the CORAL software, which is consistent with the criteria above introduced. In this chapter, we present some examples, to discuss the use of CORAL and the results which can be obtained. The following statistical characteristics were utilized in this work: (i) n is the number of compounds in a set; (ii) r^2 is the determination coefficient; (iii) q^2 is the cross

validated r^2 ; (iv) the s is root-mean squared error; (v) \overline{R}_m^2 (should be larger than 0.5) and ΔR_m^2 (should be less than 0.2) Ojha et al. (2011) as suggested metrics of predictability.

3.1 Fish Toxicity

The QSAR model for the fish toxicity was developed base on the experimental data provided in the previous works (Russom et al. 1997; Toropova et al. 2012). The data was divided into two groups—training and validation sets. This data splitting was randomly performed three times. Table 4 contains the statistical characteristics of models for fish toxicity observed in the cases of three different splits of data into the training and validation sets. The validation set is invisible (not used) during building up the model.

Table 4 The statistical characteristics of QSAR for fish toxicity obtained by (i) balance of correlations (BC); and (ii) classic scheme (CS)

Split	Method	Set	n	r^2	q^2	s	\overline{R}_m^2	ΔR_m^2
1. Eq. (5)	BC	Training	153	0.8052	0.8007	0.618		
		Invisible training	153	0.8050	0.7989	0.625		
		Calibration	130	0.8017	0.7953	0.611	0.65	0.12
		Validation	131	0.8037		0.719		
	CS	Training	306	0.8256	0.8232	0.583		
		Calibration	130	0.7593	0.7512	0.669	0.65	0.12
Validation		131	0.7839		0.739			
2. Eq. (6)	BC	Training	156	0.7804	0.7745	0.687		
		Invisible training	158	0.7677	0.7625	0.697		
		Calibration	127	0.7244	0.7152	0.625	0.55	0.13
		Validation	126	0.8065		0.663		
	CS	Training	314	0.8186	0.8163	0.613		
		Calibration	127	0.7205	0.7116	0.634	0.56	0.12
Validation		126	0.7771		0.739			
3. Eq. (7)	BC	Training	157	0.7973	0.7914	0.662		
		Invisible training	153	0.7962	0.7907	0.603		
		Calibration	129	0.7465	0.7319	0.708	0.63	0.12
		Validation	128	0.8051		0.638		
	CS	Training	310	0.8133	0.8106	0.607		
		Calibration	129	0.7475	0.7339	0.696	0.65	0.11
Validation		128	0.7942		0.640			

The fish toxicity was also investigated before the study reported here using different QSAR models. A previous model (Toropova et al. 2012) is characterized by the average correlation coefficient between pLC50 (experiment) and pLC50 (calculated) equal to 0.787. Thus, the values reported in Table 4 show that the model calculated with Eq. (2) is better than the model suggested in work (Toropova et al. 2012).

The description of models obtained for three random split is given below:

$$\text{pLC50} = -0.6459022 + 0.1023858^* \text{DCW}(1, 11) \quad (5)$$

$$\text{pLC50} = -1.0325851 + 0.0679094^* \text{DCW}(1, 6) \quad (6)$$

$$\text{pLC50} = -0.8779432 + 0.0857614^* \text{DCW}(1, 7) \quad (7)$$

3.2 Rat Toxicity

Also for rat toxicity we obtained results better than those previously published. Table 5 contains the statistical characteristics of the models for rat toxicity obtained in the cases of three splits into the training and validation sets. The validation set is invisible during building up the model. The previous model (Toropova et al. 2015a) is characterized by the average correlation coefficient between pLD50 (experiment) and pLD50 (calculated) equal to 0.754. Obviously, the predictive potential of model calculated with Eq. (3) is better than predictive potential of model suggested in work (Toropova et al. 2015a).

The models obtained for three random split are the followings:

$$\text{PLD50} = -3.1121700 + 0.0947844^* \text{DCW}(1, 12) \quad (8)$$

$$\text{PLD50} = -3.1311260 + 0.0686280^* \text{DCW}(1, 10) \quad (9)$$

$$\text{PLD50} = -3.0450211 + 0.0659596^* \text{DCW}(1, 12) \quad (10)$$

3.3 Bird Toxicity

Two models were developed for quail toxicity. Table 6 contains their statistical characteristics. The results were obtained for three data splits into the training and validation sets. The validation set is invisible during building up the model. The obtained results can be compared to the results of a previous study (Toropov and

Table 5 The statistical characteristics of QSAR for rat toxicity obtained by (i) balance of correlations (BC); and (ii) classic scheme (CS)

Split	Method	Set	n	r ²	q ²	s	\overline{R}_m^2	ΔR_m^2
1. Eq. (8)	BC	Training	151	0.7729	0.7676	0.494		
		Invisible training	150	0.7721	0.7659	0.534		
		Calibration	112	0.7189	0.7049	0.525	0.61	0.18
		Validation	112	0.7188		0.598		
	CS	Training	301	0.7889	0.7863	0.489		
		Calibration	112	0.7016	0.6879	0.512	0.59	0.09
Validation		112	0.6694		0.635			
2. Eq. (9)	BC	Training	161	0.7201	0.7136	0.588		
		Invisible training	151	0.7598	0.7531	0.509		
		Calibration	105	0.6580	0.6470	0.548	0.53	0.15
		Validation	108	0.7206		0.511		
	CS	Training	312	0.7667	0.7640	0.517		
		Calibration	105	0.6207	0.6057	0.591	0.49	0.05
Validation		108	0.7779		0.492			
3. Eq. (10)	BC	Training	160	0.7484	0.7420	0.544		
		Invisible training	155	0.7488	0.7426	0.507		
		Calibration	105	0.7155	0.7053	0.579	0.57	0.24
		Validation	105	0.6438		0.532		
	CS	Training	315	0.7866	0.7840	0.483		
		Calibration	105	0.6990	0.6867	0.592	0.58	0.24
Validation		105	0.6298		0.547			

Benfenati 2007). The previously developed model (Toropov and Benfenati 2007) is characterized by the average correlation coefficient between $\log(1/C)$ (experiment) and $\log(1/C)$ (calculated) equal to 0.731. The predictive potential of the model calculated with Eq. (4) is better than predictive potential of model suggested in work (Toropova et al. 2015a).

Three models obtained as the results of random splits of data are given below:

$$\text{Log}(1/C) = -0.7120291 + 0.1055524^* \text{DCW}(1, 14) \quad (11)$$

$$\text{Log}(1/C) = -1.0170942 + 0.0546808^* \text{DCW}(1, 7) \quad (12)$$

$$\text{Log}(1/C) = -0.8511373 + 0.0376830^* \text{DCW}(1, 5) \quad (13)$$

Table 6 The statistical characteristics of QSAR for quail toxicity obtained by (i) balance of correlations (BC); and (ii) classic scheme (CS)

Split	Method	Set	n	r^2	q^2	s	\overline{R}_m^2	ΔR_m^2
1. Eq. (11)	BC	Training	41	0.9194	0.9129	0.267		
		Invisible training	37	0.9143	0.9037	0.270		
		Calibration	18	0.9001	0.8659	0.380	0.83	0.16
		Validation	18	0.8175		0.617		
	CS	Training	78	0.9249	0.9212	0.250		
		Calibration	18	0.8183	0.7707	0.464	0.81	0.00
		Validation	18	0.7496		0.622		
2. Eq. (12)	BC	Training	38	0.7999	0.7768	0.378		
		Invisible training	41	0.8140	0.7969	0.532		
		Calibration	17	0.6668	0.5864	0.467	0.51	0.06
		Validation	18	0.7504		0.669		
	CS	Training	79	0.8101	0.8013	0.418		
		Calibration	17	0.6218	0.4918	0.520	0.61	0.00
		Validation	18	0.7448		0.590		
3. Eq. (13)	BC	Training	41	0.6616	0.6372	0.547		
		Invisible training	37	0.6751	0.6248	0.620		
		Calibration	18	0.5862	0.4030	0.599	0.58	0.00
		Validation	18	0.7875		0.473		
	CS	Training	78	0.7497	0.7373	0.478		
		Calibration	18	0.5975	0.4615	0.594	0.52	0.02
		Validation	18	0.7727		0.427		

4 Conclusions

The CORAL software provides a useful tool that can be used in various research projects. Examples of its applications to environmental studies are discussed in the chapter. QSAR models were developed to investigate toxicity of chemical compounds towards fish, rat, and quail. The split of experimental data into the training and validation sets has a clear influence on the statistical quality of QSAR models. The obtained models for three endpoints are quite satisfactory for various distributions of the data into the training and validation sets. The hybrid optimal descriptors (calculated with SMILES together with graphs) which are modifications of previously studied versions of the optimal descriptors calculated with solely SMILES provide better predictive potential in comparison with previous models suggested in the works (Toropova et al. 2012, 2015a; Toropov and Benfenati 2007) for fish, rats, and birds, respectively. As a rule, the balance of correlations approach improves the models, in comparison with the classic scheme where the training set is combination of training-calibration sets. However, some exceptions have been noticed (for instance, the second split in Table 4 for rat toxicity).

Models discussed in this chapter were built up and validated according to the OECD principles (OECD 2007). It is expected that QSAR models will be more and more applied within a broad series of studies related to environmental endpoints, and the recent review by Roy and Kar (2016) is a clear example of the high number of studies in this sector. The collections of data, as in this review, could be easily used to open out the CORAL software for further applications.

Acknowledgements The authors are grateful for the contribution of the EC project LIFE-COMBASE (LIFE15 ENV/ES/000416). This work was financially supported by National Science Foundation: NSF-CREST grant #HRD-1547754.

References

- OECD. (2007). (Organization for Economic Co-operation and Development). Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] Models No. 69.
- Ojha, P. K., Mitra, I., Das, R. N., & Roy, K. (2011). Further exploring r_m^2 metrics for validation of QSPR models. *Chemometrics and Intelligent Laboratory Systems*, 107, 194–205.
- Roy, K., & Kar, S. (2016). In Silico models for ecotoxicity of pharmaceuticals. In: E. Benfenati (Ed.), *Silico methods for predicting drug toxicity*. Methods in molecular biology 1425 (pp. 237–304). Springer.
- Russom, C. L., Bradbury, S. P., Broderius, S. J., Hammermeister, D. E., & Drummond, R. A. (1997). Predicting modes of action from chemical structure: acute toxicity in the Fathead minnow (*PimephalesPromelas*). *Environmental Toxicology and Chemistry*, 16, 948–957.
- Toropov, A. A., & Toropova, A. P. (2002a). Modeling of acyclic carbonyl compounds normal boiling points by correlation weighting of nearest neighboring codes. *Journal of Molecular Structure: THEOCHEM*, 581, 11–15.
- Toropov, A. A., & Toropova, A. P. (2002b). QSAR modeling of toxicity on optimization of correlation weights of Morgan extended connectivity. *Journal of Molecular Structure: THEOCHEM*, 578, 129–134.
- Toropov, A. A., & Benfenati, E. (2007). Optimisation of correlation weights of SMILES invariants for modelling oral. *European Journal of Medicinal Chemistry*, 42, 606–613.
- Toropov, A. A., Toropova, A. P., Martyanov, S. E., Benfenati, E., Gini, G., Leszczynska, D., et al. (2011). Comparison of SMILES and molecular graphs as the representation of the molecular structure for QSAR analysis for mutagenic potential of polyaromatic amines. *Chemometrics and Intelligent Laboratory Systems*, 109, 94–100.
- Toropov, A. A., Toropova, A. P., Puzyn, T., Benfenati, E., Gini, G., Leszczynska, D., et al. (2013). QSAR as a random event: Modeling of nanoparticles uptake in PaCa2 cancer cells. *Chemosphere*, 92, 31–37.
- Toropov, A. A., Toropova, A. P., Benfenati, E., & Fanelli, R. (2016). QSAR as a random event: Selecting of the molecular structure for potential anti-tuberculosis agents. *Anti-Infective Agents*, 14, 3–10.
- Toropova, A. P., Toropov, A. A., Lombardo, A., Roncaglioni, A., Benfenati, E., & Gini, G. (2012). CORAL: QSAR model for acute toxicity in Fathead Minnow (*Pimephalespromelas*). *Journal of Computational Chemistry*, 33, 1218–1223.
- Toropova, A. P., Toropov, A. A., Benfenati, E., Leszczynska, D., & Leszczynski, J. (2015a). QSAR model as a random event: A case of. *Bioorganic & Medicinal Chemistry*, 23, 1223–1230.

- Toropova, A. P., Toropov, A. A., Veselinović, J. B., & Veselinović, A. M. (2015b). QSAR as a random event: a case of NOAEL. *Environmental Science and Pollution Research International*, 22, 8264–8271.
- Toropova, A. P., Schultz, T. W., & Toropov, A. A. (2016). Building up a QSAR model for toxicity towards *TetrahymenaPyriiformis* by the Monte Carlo method: A case of benzene derivatives. *Environmental Toxicology and Pharmacology*, 42, 135–145.
- Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28, 31–36.
- Weininger, D., Weininger, A., & Weininger, J. L. (1989). SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences*, 29, 97–101.
- Weininger, D. (1990). SMILES. 3. Depict. Graphical depiction of chemical structures. *Journal of Chemical Information and Computer Sciences*, 30, 237–243.

Environmental Toxicity of Pesticides, and Its Modeling by QSAR Approaches

Mabrouk Hamadache, Abdeltif Amrane, Othmane Benkortbi,
Salah Hanini, Latifa Khaouane and Cherif Si Moussa

Abstract Thousands of environmental pollutants including pesticides, issued from human activities, are accumulated in the environment making a source of danger for the whole ecosystem. Also, the risk assessment process has become a vital and necessary discipline in the legislation to ensure that these pollutants pose no risk or negligible risk to human health, wildlife and the whole ecosystem. The risk assessment carried out for the three natural compartments, namely the terrestrial, the aquatic environment and air, is usually based on experimental studies whose cost is especially high in terms of money, time and laboratory animals. Thus, regulatory agencies are turning to the search for alternative methods less expensive, reliable and fast, which may have a power to predict the potential risks of chemical pollutants. One such toxicological predictive approach is obtained by the development of quantitative models of structure-activity relationships (QSAR). They provide the means for estimating the toxicity of a variety of chemicals in the absence of experimental data on toxicity. In this chapter, a review of publications dedicated to

M. Hamadache (✉) · O. Benkortbi · S. Hanini · L. Khaouane · C. Si Moussa
Laboratoire des Biomatériaux et Phénomènes de Transport (LBMPT),
Université de Médéa, Quartier Ain D'heb, 26000 Medea, Algeria
e-mail: mhamdeche@yahoo.fr

O. Benkortbi
e-mail: benkortbi_oth@yahoo.fr

S. Hanini
e-mail: s_hanini2002@yahoo.fr

L. Khaouane
e-mail: latifa_khaouane@yahoo.fr

C. Si Moussa
e-mail: simoussa_cherif@yahoo.fr

A. Amrane (✉)
Ecole Nationale Supérieure de Chimie de Rennes, Université de Rennes 1,
CNRS, UMR 6226, 11 allée de Beaulieu CS 50837, 35708 Rennes Cedex 7, France
e-mail: abdelatif.amrane@univ-rennes1.fr

© Springer International Publishing AG 2017

K. Roy (ed.), *Advances in QSAR Modeling*, Challenges and Advances
in Computational Chemistry and Physics 24, DOI 10.1007/978-3-319-56850-8_13

pollution by pesticides and their effects on the entire ecosystem is described. The general principles of the development and validation of QSAR models are also described. Then a critical review of QSAR models published in the literature to date for the prediction of the toxicity of pesticides is also covered.

Keywords Toxicity • Pesticides • QSAR models • Prediction

1 Introduction

For a long time the main concern of humankind has been to ensure food security given the rapid population growth. To do this, it has been imperative to take all measures to increase agricultural production. However, this objective could not be achieved without a struggle against all organisms responsible for crop damage. The discovery of pesticides was hailed as a major breakthrough for mankind. After the synthesis of dichlorodiphenyltrichloroethane (DDT) in 1939 by Paul Müller, the number and amount of pesticides has grown continuously. Diversified pesticides were produced and pesticide consumption worldwide has increased dramatically from 1960 to 2012 (Fig. 1).

The pesticides are a diverse group of inorganic and organic chemicals widely used against insects, fungi, rodents, noxious weeds, etc. The conventional pesticides are classified as herbicides, insecticides, nematocides, and fungicides. The worldwide consumption of pesticides is estimated over 2.27 million ton each year for agricultural, residential, commercial or industrial settings (Saeedi Saravi and Dehpour 2016).

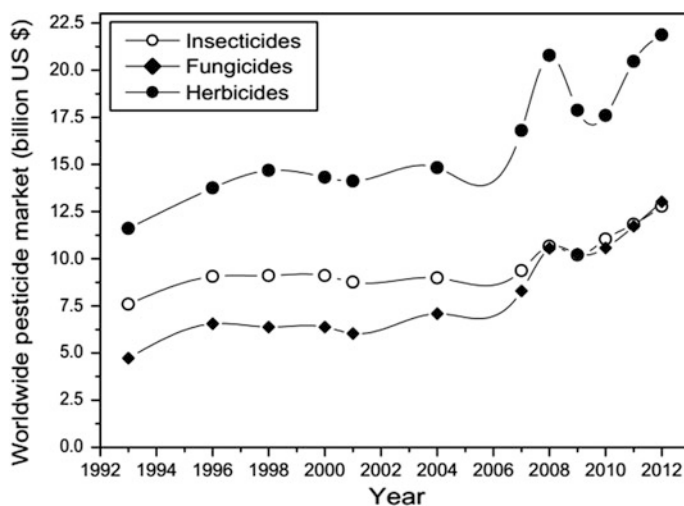


Fig. 1 Worldwide pesticide market (Peshin and Zhung 2014)

To enhance the production of foodstuffs needed to meet the needs of the ever expanding human population ensuring stable and predictable food supplies and high productivity, the pesticides are widely used in agriculture, and horticulture. They are also used in domestic applications, to slow the spread of insects, to maintain lawns, recreational areas and highways. Pesticides have also contributed to the control of many human diseases transmitted by insects. In view of the current intensification of agricultural activities and increased intensity of pesticide use, and despite their advantages, these compounds entail a number of risks and problems. Indeed, many studies made internationally highlight the environmental pollution by pesticides. They are found in the environment in all parts of the world, both in areas where pesticides are used and in areas where they never have been used. Due to the excessive use of these products, they are found in the environment (water, soil, air) and in terrestrial and aquatic food chains. In addition, they also pose a threat to the environment, humans, animals and other organisms.

The health hazards of pesticides are a major concern internationally. Long-term exposure to pesticides can cause harm to human life and can disrupt the functioning of various organs in the body. This significant relationship between exposure to pesticides and some chronic diseases has been the subject of several scientific publications. Acute poisonings by agricultural pesticides are currently considered to be an important cause of human morbidity and mortality worldwide, with some 26 million human pesticide poisonings and with about 220,000 deaths per annum in the world (Peshin and Zhang 2014). Furthermore, the discovery of pesticide residues in various sections of the environment has raised serious concerns. As a result, human beings are exposed to the effects of these compounds by eating foods in contact with contaminated soil or water.

As seen, humans and the environment are exposed to thousands of pesticides. This pollution caused by pesticides has become an important issue affecting the survival and development of human being. It is evident that risk assessment for pesticides can provide a precaution against the corresponding pollution. One of the procedures currently used for human and environmental risk assessment is the determination of the acute toxicity of pesticides. Toxicity studies aim at investigating the effects of pesticides in laboratory animals exposed to various dosage regimen for different durations. The information from toxicity studies is used in hazard and risk assessment of pesticides occurring in foodstuffs, in water, and in air. Unfortunately, experimental determination of the toxicity takes time, requires a high expense and poses an ethical problem (demands to reduce or abolish the use of animals). Also, there is a very large body of research going on in many countries with the aim of replacing *in vivo* tests by *in silico* prediction methods according to the European Directive on the Protection of Laboratory Animals (Golbamaki et al. 2014) and the Registration, Evaluation, Authorization and restriction of Chemicals (REACH) regulation (Cassotti et al. 2014). Despite being significantly cheaper than *in vivo* study, *in vitro* tests are still costly compared with *in silico* methods (Sazonovas et al. 2010). The use of *in silico* predictive methods, based on computer

tools, offers a rapid, cost-effective and ethical alternative to testing toxicity of chemical substances in animals (Sullivan et al. 2014). These methods include the Quantitative Structure–Activity Relationship (QSARs) models. QSAR models describe a mathematical relationship between the structural properties of a set of compounds and the particular activity (toxicological or other), associated with them. The use of QSAR in environmental studies experienced an increasing development in the past two decades. One of the main areas of interest in these studies is the modeling and prediction of toxicological effects.

The aim of this chapter is to briefly review in the literature, the QSAR models established for predicting the toxicological properties of pesticides. In Sect. 2 we will talk about pesticide pollution of all components of our environment and that of food. In Sect. 3 we give an overview of proven or suspected impacts of pesticides on health. Section 4 briefly presents the QSAR models for predicting the toxicological properties of pesticides published in literature so far. In Sect. 5 some conclusions are presented.

2 Pesticides and Pollution

Depending on the conditions of use and the characteristics of the environment, pesticides are likely to be found in different compartments of the environment and in food. For nearly fifty years, pesticides have been detected in the waters of rivers and groundwater, air and rainwater. They are also found in fruits, vegetables, cereals and animal products (eggs, milk, meat, fish, etc.). They exist in their original form, but they can also be degraded (residues or metabolites). The potential of a pesticide to move depends on its chemical properties (ionization, water solubility, volatility, persistence in the environment), its formulation, soil properties (moisture content, pH, percentage of organic matter), the rate and method of application, weather conditions (frequency and distribution of rainfall) and the depth of the water. Other methods which influence the fate of the chemical include the absorption of the plant, the adherence of soil, and the volatilization.

Many works made internationally highlight the environmental pollution by pesticides. The literature review in this area has focused on the pollution of the various environmental compartments. The literature review was carried out on the basis of keywords in PubMed using combinations of the following keywords: ‘soil pollution by pesticides* air pollution by pesticides* water pollution by pesticides* foodstuffs pollution by pesticides*’ in Topics. We retrieved several thousands publications (Fig. 2). Those which appeared relevant for the review were sorted using the titles, the abstracts and the full texts. To complete the review, starting from the selected references, authors contributing to the references on the subject of interest were identified and all their publications were studied.

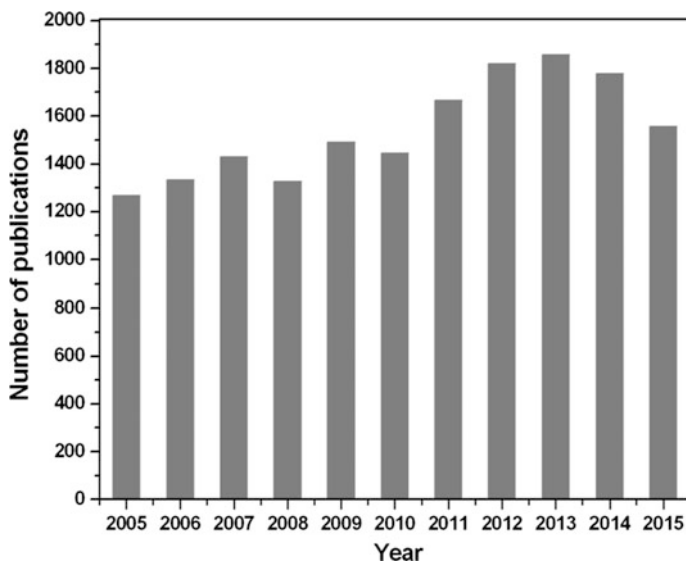


Fig. 2 Number of publications on the environmental contamination by pesticides (PubMed accessed 18/03/2016)

2.1 Pesticides and Environmental Pollution

2.1.1 Atmospheric Pollution

Pesticides in the atmosphere are mainly from the release of pesticides from treated plants, evaporation of pesticide residuals in soils and water bodies, and volatilization of pesticides sprayed. This contamination is chronic.

Studies from various research groups (Briand et al. 2003; Scheyer et al. 2005; Takazawa et al. 2016; Yusà et al. 2014; Hogarh et al. 2013; MOEJ 2015; Schummer et al. 2010; Borrás et al. 2011; Coscolla et al. 2013) conducted pesticide concentration measurements in the atmosphere. All these works revealed the presence of pesticides in all atmospheric phases, whether gaseous, liquid or particulate in aerosols, droplets fog or rain (Bedos et al. 2002). In another study, researchers were able to observe in rainwater in Denmark a number of compounds banned in that country but permitted in other European countries, indicating a significant contribution of atmospheric transport in the local contamination (Asman et al. 2005). An active substance (Epoxiconazole fungicide) used on plants was found in the air (Coscolla et al. 2010), despite its low volatilization. A similar observation was made for the insecticide chlorpyrifos, which was also found in the atmosphere (Yao et al. 2008; Zhou et al. 2010).

A recent study (Gunier et al. 2011) shows that agricultural pesticides used near homes eventually contaminate the inside of these houses. This study shows that the use of agricultural pesticides, some of which are suspected carcinogens, has an

impact on the contamination of the air breathed every day by their inhabitants. Furthermore, the analysis during a full year of air of Yangtze (China) revealed the presence of high concentrations of organochlorine pesticides. The authors claim that air pollution may cause cancer by inhalation following the results of the calculated risk factor (Zhang et al. 2013). In another study, 6 pesticides were analyzed in the atmosphere of northern Algeria by Moussaoui et al. (2012). Pesticides found in urban gas phase showed high levels relative to sampling a rural area. Malathion and chlorpyrifos were present at high concentrations.

In a study conducted in the atmosphere of Spain, Hart et al. (2012) have detected 24 pesticides with different concentrations. The same authors also have reported the levels of 17 more polar pesticides in the Valencia Region in Spain (Coscolla et al. 2013). More recently Raepfel et al. (2014) have described the levels of 9 pesticides detected in the atmosphere of Strasbourg (France).

2.1.2 Water Pollution

This part describes information with respect to pesticide levels found in aquatic environments. Water contamination by pesticides has become increasingly worrisome. Given the risks they represent, the presence of pesticides in rivers, coastal waters and groundwater is the subject of regular monitoring that has steadily increased during the last decade. The concentration of pesticides in streams has been reported by several authors. Several recent research studies have shown that in many parts of the world, water systems show significant contamination by pesticides (Cruzeiro et al. 2016; Liu et al. 2016a; Haddaoui et al. 2016; Wu et al. 2014; Dabrowski et al. 2014; Rasmussen et al. 2015; De Geronimo et al. 2014; Silva et al. 2015; Palma et al. 2014; Papadakis et al. 2015; Grung et al. 2015; Zheng et al. 2016). The levels of waters' pollution by pesticides are different and can be ranked as: cropland water > runoff > pond water > groundwater > river water > deep groundwater > sea water (Zhang et al. 2011).

For example, the small coastal rivers (Louros and Arachthos) in Greece have occasionally substantial herbicide content (Munaron 2004). In addition, Steen et al. (2001) detected on the Scheldt River, peaks of residues of certain pesticides far in excess of the required standard. In aquatic compartments, traces of chlorpyrifos were also detected (Coupe and Blomquist 2004). In Poland, studies indicate contamination of precipitation by pesticides (Polkowska et al. 2000; Gryniewicz et al. 2001). In studies on contamination of waters of the Aquitaine region of France led by Barjhoux (2011), it was found that 2.6% of contaminants are pesticides.

Surface water samples were collected in Mississippi to assess pesticide levels in two separate studies. Traces of pesticides as metolachlor and acetochlor and some of their metabolites have been reported by Rebich et al. (2004). For their part, Tagert et al. (2014) detected hexazinone in 94% of samples, followed by metolachlor (76%), tebuthiuron (48%), atrazine (47%) and metribuzin (6%). Italian water resources are contaminated with terbuthylazine and its metabolite (Desethyl terbuthylazine); this was reported by Bottoni et al. (2013) in a study of this potentially

toxic pesticide. In Guadeloupe, organochlorine pesticide residues were detected in very high concentrations in freshwater ecosystems. These levels far exceed the permitted limit of residual Chlordecone in fish and shrimp (Coat et al. 2011).

Moreover, the majority of rivers and sources of drinking water in India are contaminated by pesticides (Agrawal et al. 2010). In Pakistan, a number of samples of fish and shellfish were analyzed for the determination of pesticide contamination due to pollution of the marine environment. For the authors, the results of this study do not seem to be very alarming compared to trace levels found in other parts of the world (Hina et al. 2013). An analysis of the Guanting reservoir waters (China) revealed the presence of no less than 18 kinds of organochlorine pesticides (Wan et al. 2009). A total of 27 samples of shallow groundwater were collected from the Taihu Lake region in China. DDT and hexachlorocyclohexane (HCH) are the most predominant contaminants in those waters. In a very recent study on the contamination of shallow groundwater, Wu et al. (2014) concluded that the calculated values of the carcinogenic risk of contaminants raise a risk of potentially serious cancer for those who consume drinking water from these waters.

2.1.3 Soil Pollution

According to various studies conducted all over the world, it was established that a variety of soils, including cultivated fields, vegetable fields and forest land are contaminated by various pesticides (Sun et al. 2016a, b; Qu et al. 2015; Gao et al. 2013; Zhang et al. 2006; Verma et al. 2014). The best documented examples relate organochlorine pesticides (Galiulin et al. 2002). For example, in a monitoring carried out in 2002, Mast et al. (2003) estimated that the annual input to the soil level reaches 45.8, 14.2 and 54.8 mg/ha for atrazine, dacthal and carbaryl, respectively.

The effect of this contamination is decreasing soil fertility following the destruction of microorganisms and earthworms. In a study on indicators of soil contamination by pesticides, Floch et al. (2011) suggest that pesticides may indirectly affect the enzymatic activities of soil through their action on soil microorganisms. Furthermore, contamination by organochlorine pesticides of two soils located respectively near an old and new pesticide factory was the subject of a study in China (Zhang et al. 2009). This study found that these soils were contaminated differently by the following materials: HCH, DDT, HCB and chlordane. According to Zhao et al. (2009), different types of soil in the area of Haihe River (China) are contaminated with pesticides DDT and HCH. The same pesticides were also found in the Pearl River Delta soils in China (Ma and Ran 2009).

Degradation of pesticides according to the nature of the soil has been the subject of studies. For example, in a study of the nature of the soils that suffer most from contamination by chlorpyrifos pesticide, Chai et al. (2013) conclude that the degradation of this substance is slower for acid soils, soils with high clay content and soils at low temperature. A study by Oukali-Haouchine et al. (2013) showed that metribuzin is effectively adsorbed by Algerian clay soils. However, the

adsorbed amounts remain low. According to the authors, about 3/4 of the metribuzin introduced are not retained by the soil and could be transferred into the groundwater, which could pose a significant risk of groundwater contamination.

2.2 *Pesticides and Food Contamination*

Pesticides continue to be used in the production of foods. They are therefore dangerous substances for living beings and humans. Pollution is caused by spraying pesticides, seed treatment and soil treatment with pesticides. Pesticide residues are found in agricultural and animal products such as wheat, corn, fruits, vegetables, cereals, tea, fish, milk, eggs, meat, honey and medicinal herbs (Lozowicka et al. 2014; Shoiful et al. 2013; Tsakiris et al. 2015; Wu et al. 2013a; Feng et al. 2015; Skretteberg et al. 2015; Arias et al. 2014; Xu et al. 2015; Calatayud-Vernich et al. 2016; Barjanska et al. 2013; Juan-Borras et al. 2016; Liu et al. 2016b; Yuan et al. 2014; Wang et al. 2014; Ahsan et al. 2013; Singh et al. 2014).

In a study on the contamination of crops on land contaminated with chlordane, it was noticed a transfer of this pesticide to crops. All tubers of different crops (sweet potato, turnip, radish, zucchini and tomato) were contaminated (Cabidoche and Lesueur-Jannoyer 2012). Nougadère et al. (2012) have investigated the presence of residues of three hundred twenty five pesticides and their transformation products in food samples covering 90% of the diet of a population. The results showed that 37% contained one or more residues. Seventy-three pesticides were detected and quantified. The most frequently detected pesticides were insecticides pirimiphos-methyl and chlorpyrifos-methyl. Dimethoate pesticide and its metabolite were detected in two samples of cherries at levels above the allowable daily dose. A study on phosphorylated pesticide residues in the Kuwait's food was undertaken by Saeed et al. (2005). The results indicated that 18% of the samples contained residues. Monocrotophos (0.2 mg/kg), diazinon (0.05 mg/kg), quinalphos (0.022 mg/kg), chlorpyrifos-methyl (0.01–0.33 mg/kg) and fenitrothion (0.16–0.84 mg/kg) were the most frequently detected pesticides. The use of pesticides in vegetable production was the subject of a study in the rural town of Tori-Bossito in southern Benin (Ahouangninou et al. 2012). Pesticide residues were found in 42% of samples of eggplant leaves, cucumber, amaranth and Solanum. In Ghana, a study was conducted to evaluate the residues of organochlorine and organophosphorus pesticides in fruits and vegetables sold in markets (Bempah et al. 2012). 9.8% of 309 samples of fruits and vegetables showed rates of residues above the permitted limit.

Eight different pesticides were detected in the analysis of pesticide residues in wheat imported by South Africa. The most frequently detected pesticides were mercaptothion, permethrin and chlorpyrifos. The authors point out that this wheat-based foods could be a source of contamination for both humans and animals (Dalvie and London 2009). Moreover, the pesticides diazinon, chlorpyrifos and quinalphos were found in analyzes of a variety of fruits (Sanagi et al. 2013). Other

food products including fruit, vegetables and cereals were the subject of analysis. Thus, organochlorine and pyrethroid pesticide residues were found in samples of tea during a scan through the use of a new chromatographic technique, sensitive and effective (Liu and Min 2012). The contamination of milk and milk products was the subject of a scientific publication (Fischer et al. 2011). This article gives an overview of the nature, sources, appearance, detection and the potential risk for human health of major chemical contaminants of this aliment. Traces of pesticides (DDT, lindane, dieldrin, etc.) have actually been detected. Farajzadeh et al. (2014) have highlighted a chromatographic method that allowed them to observe the contamination of vegetable oils by the following pesticides: fenpropathrin, Sumithrin, cyhalothrin, permethrin and deltamethrin. We can also report the existence of a study whose results showed that the concentrations of organochlorine pesticides are higher in honey from developing countries than in honey of developed countries (Wang et al. 2010).

Davodi et al. (2011) have determined the concentrations of pesticides in 8 fish species collected from marshes Shadegan in Iran. In all samples, the concentrations were higher than the guidance standards for food safety issued by the European Union (EU) and the Food and Drug Administration (United States). Other researchers (Arzi et al. 2011) calculated the concentrations of aldrin, dieldrin, heptachlor, heptachlor epoxide and methoxychlor in fish caught in the province of Khuzestan in Iran. All fish examined were contaminated with organochlorine pesticides studied, with concentrations found for some pesticides. Furthermore, a survey was conducted to evaluate pesticide residues in fish samples from the river Densu in Ghana (Fianko et al. 2011). The data obtained indicate that the rate of γ -HCH, heptachlor, of α -endosulfan, endosulfan sulfate and dieldrin exceeded the reference dose, thereby indicating great potential for systemic toxicity consumers. Akan et al. (2013) used four fish species of Borno, one of Nigeria's states, for residue analysis. Eleven organochlorine pesticides were detected in all the samples examined. This study also revealed that all pesticide residues in samples of fish studied were above the maximum allowable limits.

3 Actual or Suspected Health Impacts of Pesticides

Although pesticides have largely benefited mankind through the development of agricultural products and the control of infectious diseases, their intensive use, in turn, threatens human health and environmental components. The long-term exposure to pesticides can harm human life and can disrupt the functioning of various organs in the body, including the nervous, endocrine, immune, reproductive, renal, and cardiovascular and respiratory systems (McKinlay et al. 2008; Jiang et al. 2011; Ali et al. 2014; Cachot 2014; Koureas et al. 2012; Ge et al. 2013; Lee et al. 2015; Saedi Saravi and Dehpour 2016; Lebov et al. 2016; Zhang et al. 2016). In this regard, there is evidence on the link between pesticide exposure and the incidence of human chronic diseases, such as cancer, Parkinson's disease,

Alzheimer's disease, diabetes, aging, cardiovascular disease and chronic kidney disease (Mostafalou and Abdollahi 2013; Van Maele-Fabry et al. 2011, 2012, 2013; Furlong et al. 2015; Jaacks and Staimez 2015; Evangelou et al. 2016; Lerro et al. 2015).

This significant relationship between exposure to pesticides and some chronic diseases has been the subject of several scientific publications. The literature review was carried out on the basis of keywords in PubMed, with the following formula: 'health effect* and (pesticide* or herbicide* or fungicide* or insecticide*)' in Topics. We retrieved more than 3200 publications between 2005 and 2015. Those which appeared relevant for the review were sorted using the titles, the abstracts and the full texts. To complete the review starting from the previously selected references, authors contributing to the papers on the subject of interest were identified and their publications were studied. This allowed us to select relevant references. By way of an illustration, is shown below in Fig. 3 the number of scientific publications of the years 2005–2015.

3.1 Pesticides and Reproductive Disorders

For reproductive disorders, studies have suggested the possibility of a link between pesticide exposure and the risk of male infertility, excess of spontaneous abortion, premature, stillborn and certain fetal malformations (AIRPARIF 2007). Research

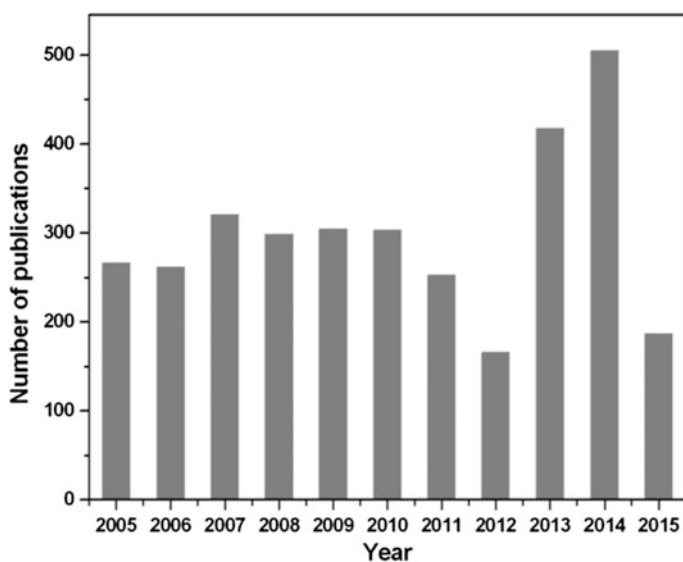


Fig. 3 Number of publications inherent to the effects of pesticides on health (PubMed accessed 03/22/2016)

showed that conazole fungicides (e.g., Epoxiconazole) act as endocrine disruptors (Kjaerstad et al. 2010). Also, it was observed in rats a disruption of reproductive development such as longer duration of gestation or fetal death (Christiansen et al. 2010) and in quail, a reduction of the number of spermatids (Grote et al. 2008).

A wide range of literature details the adverse effects of environmental exposures, including pesticides, and on both male and female reproductive systems (Shojaei Saadi and Abdollahi 2012). Following a study on the evaluation of reproductive disorders in men and women working in greenhouses, Bretveld et al. (2008) claimed to have evidence on the assumption that pesticide exposure affects human reproduction leading to miscarriage. The general population may be affected by these fertility problems due to exposure to pesticides. Thus, another study (Orton et al. 2011) reveals that pesticides to which the European population is exposed are potentially threatening to the male fertility.

A study by Andersen et al. (2008) shows that boys whose mothers worked in greenhouses where pesticides were used during pregnancy have a development of reproductive functions disrupted. The study focused on the development of reproductive functions of 110 boys. The results of the study demonstrate the prevalence of cryptorchidism (undescended testicles), a lower testicular volume and a lower serum concentration of testosterone in boys of mothers exposed to pesticides in greenhouses than boys of not exposed mothers. For the authors, these results suggest a negative effect of the professional use of pesticides by mothers, during pregnancy, on the development of function reproduction of their boys, despite the precautions taken for use.

3.2 Malformations and Immune System Disorders

A recent study (Chevrier et al. 2011) reinforces the hypothesis that environmental exposure of pregnant women to the pesticide atrazine increases the risk of adverse effects on the fetus. Scientists followed a cohort of pregnant women from 2002 to 2006 in Britain. The authors noted increased risks of low weight and small head circumference at birth for babies whose mothers were exposed to atrazine by their environment. In mammals, it has been observed that insecticides affect the immune system (Nandi et al. 2011).

3.3 Pesticides and Cancer

The pesticides DDT, lindane, simazine and the risk of prostate cancer occurrence were mentioned in a study on the exposure of farmers of British Columbia to these compounds (Band et al. 2011). In later studies on the impact of pesticides on the

health of agricultural workers, Mills and Yang (2007) arrived at a similar conclusion to that of other studies on gastric cancer conducted in Europe. They suggest that the risk of stomach cancer experienced by agricultural workers can be attributed to exposure in the workplace. Chrisman et al. (2009) evaluated statistically the degree of correlation between pesticide sales in 1985 and the death rate from cancer between 1996 and 1998 in 11 Brazilian states. The results suggest a significant correlation with a coefficient varying from 0.61 (esophageal cancer) to 0.73 (lip cancer). In addition, these authors suggest that the population exposure to pesticides in the 1980s in some Brazilian states may be associated with the development of cancer a decade later.

Provost et al. (2007) showed an increased risk of developing certain brain cancer for people exposed to pesticides through their professional activities or home. The study found that the increased risk is statistically significant for strong pesticide exposure levels. For farmers exposed to the higher levels, the risk is more than doubled. Women working on the farm are more likely to develop breast cancer than others, according to a Canadian study (Brophy et al. 2002). Researchers who have studied the work history of 564 women with breast cancer in the Windsor area (central Canada), noted that the risk was 2.8 times higher for those who worked in a farm a moment of their lives.

The risk of leukemia in children is associated with maternal exposure to pesticides during pregnancy (Wigle et al. 2009). The researchers studied the results of 31 epidemiological studies published between 1950 and 2009 studying the link between childhood leukemia and pesticides exposure for parents. The result states that the leukemia risk was doubled among children whose mothers were occupationally exposed to pesticides during pregnancy, compared to children of unexposed women. For farmers, the risk was increased by 40%. Moreover, the risk of leukemia is related to the nature of the pesticide.

3.4 Pesticides and Diabetes

Users of pesticides who used chlorinated pesticides for more than 100 days in their lifetime have an increased risk of diabetes, according to a study from the National Institute of researchers for the US Health (Montgomery et al. 2008). Depending on the nature of pesticides, the risk may be increased by 20–200%. This study which was conducted on more than 30,000 farmers shows that among the 50 different pesticides that the researchers studied, 7 products especially have retained their attention (aldrin, chlordane, heptachlor, dichlorvos, trichlorfon, alachlor and cynazine). More Recently several studies confirmed that some pesticides are significantly associated with type 2 diabetes (Wu et al. 2013b) and with abnormal glucose tolerance determined from oral glucose tolerance test among obese individuals (Dirinck et al. 2014).

3.5 Pesticides and Neurological Pathologies

Among the delayed neurotoxic effects that may be related to the use of pesticides, Baldi et al. (2012) hold the following disorders: polyneuropathy, neuropsychological disorders, Parkinson's disease. In another study on the delayed effects of pesticides on human health, it was concluded that occupational exposure to organochlorine and organophosphate pesticides is associated with the onset of neuropsychological and neurobehavioral disorders (Multigner 2005). Several studies showed that people exposed to pesticides (insecticides and herbicides) are those at high risk of contracting Parkinson's disease (Freire and Koifman 2012; Van Maele-Fabry et al. 2012; Van der Mark et al. 2012). During tests on pesticides, some authors have found a significant direct link between occupational exposure to organophosphate pesticides and the development of Alzheimer's disease in their lifetime (Hayden et al. 2010). Moreover, Parron et al. (2011) showed that people living in areas with high levels of pesticide have a high risk of contracting Alzheimer's disease. Exposure to pesticides also seems linked to a greater risk of developing Parkinson's disease. Another recent study (Costello et al. 2009) shows that exposure to pesticides maneb and/or paraquat increases on average by 75% the risk of developing Parkinson's disease in people exposed. The risk is multiplied by 2.27 following exposure to one of these two pesticides or multiplied by 4.17 when young subjects are exposed to the two pesticides.

3.6 Respiratory Affection and Pesticides

An analysis (Hoppin et al. 2007) showed an excess risk for the occurrence of chronic bronchitis during the use of two insecticides, diazinon and Malathion. Furthermore, in a study on farmers employing forty pesticides, wheezing's respiratory were reported by 19% of farmers (Hoppin et al. 2002).

4 Quantitative Structure-Activity Relationships (QSAR)

4.1 Introduction

Thousands of pesticides are released into the environment. Several studies have shown the toxic potential of these compounds on human health and wildlife (Mas et al. 2010). Also, regulators worldwide look for assessing toxicological and ecotoxicological risks posed by the release of these substances. The risk assessment process is traditionally performed using experiments on laboratory animals. These tests represent 8% of the total number of animals used in experiments (Devillers and Devillers 2009). However, a full evaluation of the toxic properties by experimental testing is time consuming, expensive and poses an ethical problem.

Given the growing demand for toxicity assessment, many countries and/or organizations are studying the use of alternatives to animal testing. Such is the case of REACH (Registration, Evaluation and Authorisation of CHemicals), the new European legislation on chemicals, which requires, when it is possible, to replace the animal tests by alternative methods (Devillers and Devillers 2009). The aim of REACH is to improve the protection of human beings and the environment through a better and earlier identification of the toxic properties of the compounds (EU 2006). We can also mention FRAME (Fund for the Replacement of Animals in Medical Experiments: <http://www.frame.org.uk>) which is engaged in the development and acceptance of methods to replace animal testing for regulatory and other purposes (Dearden 2002).

The use of computer technology to predict the chemical impact on the environment and human health is an alternative to animal testing. With software tools, several toxicity prediction approaches were developed. Among these approaches were the traditional QSAR methods, qualitative structure-activity relationships (SAR) methods, expert systems and 3D-QSAR like comparative molecular field analysis (CoMFA). The advantage of these prediction methods is their low cost, short duration, high efficiency and reproducibility using the same model. In this chapter, we will focus on QSAR methods for predicting the toxicity of pesticides.

4.2 The General Principles of QSAR Models for Toxicity Prediction

The QSAR Toxicological models are mainly used for predicting the toxicity of new compounds. These models are mainly developed from the training set of compounds with known activity. At the origin of any model, there is a basic assumption. In the case of QSAR for toxicology, it is assumed that the toxicity is related to the chemical structure. Also, a QSAR is a mathematical equation that correlates a particular activity (toxicological or other) and the structural properties of a series of compounds. This equation can then be used to predict toxicity or property of other compounds. The use of QSAR models is of great importance in the risk assessment of pollutants because they provide quick answers, reliable and quite accurate. They can serve as alternatives to existing experimental techniques. It is useful to note that the commercial models that can predict the toxicity of chemicals (such TOPKAT, CAESAR or DEREK) are available on the market (Venkatapathy and Wang 2013).

To be used in a regulatory context, the QSAR models must respect the five principles proposed by the Organisation for Economic Co-operation and Development (OECD) (Dearden and Rowe 2015): (1) a defined endpoint (2) an unambiguous algorithm, (3) a defined domain of applicability, (4) appropriate measures of goodness of fit, robustness, and predictivity, (5) a mechanistic interpretation, if possible. In particular, a series of internal and external validation tests are used to demonstrate the reliability of a QSAR model. The goodness of fit is evaluated for

the compounds used to establish the model (i.e., the compounds of the training set). Robustness is estimated by cross-validation and/or randomization techniques. The predictive power is generally evaluated by means of a set of external validation of compounds which are not used to develop the model. Currently, external validation of QSAR models is a standard requirement.

The construction of a QSAR model for toxicological prediction requires the following three phases:

1. An experimental base of high quality for the studied chemical compounds. It must include the following elements: structure of compounds, physico-chemical and toxicological properties with reference to the species, strain and laboratory animal sex used (Zakharov and Lagunin 2014). Note that a huge amount of information from toxicity databases is free. For pesticides, we can cite two databases: the Pesticide Properties DataBase (PPDB) and the Pesticide Action Network (PAN). However, the appropriate selection of high quality experimental data, which will be used to create the model, is of utmost importance for the development of a reliable model.
2. In the second phase, it is necessary to calculate the descriptors. The latter are used for the description of chemical structures. There are software programs that generate different types of molecular descriptors. One of the most popular software provided by Talete Italian company is called Dragon. In addition, several methods are used to select the most relevant descriptors. The regression, principal component analysis (PCA) and genetic algorithms (GA) have been proposed in the literature. These are probably the most powerful tools because they are able to fully explore the molecular space (Lagunin et al. 2011).
3. The next step is to use a mathematical method to identify the relationship between the descriptors and their biological effects (e.g., toxicity). There are several manuals with descriptions of mathematical methods (or statistical learning techniques) used in modeling QSAR (Zakharov and Lagunin 2014). The best known methods are: linear regression (LR), multilinear regression (MLR), and nonlinear techniques such as artificial neural networks (ANN) and support vector machines (SVM).

4.3 QSARs for Predicting Toxicity of Pesticides

The REACH legislation (Dearden and Rowe 2015) recommends evaluation of nineteen toxicological properties in terms of annual tonnage. Among these properties, we can cite: skin and eye irritation, acute toxicity, aquatic toxicity, reproductive toxicity, chronic toxicity, effects on terrestrial organisms, the carcinogenic effect and long-term toxicity for invertebrates, micro-organisms and sediments, plants and birds. In the present section, we try to examine and evaluate QSAR models published in the literature devoted to the prediction of the toxicological properties.

4.3.1 QSARs for Acute Toxicity

Acute toxicity describes the adverse effects caused by a single exposure to a chemical substance. Exposure is generally oral, dermal or inhalation. LD₅₀ (Lethal Dose) is a way to measure the acute toxicity of a given compound. The LD₅₀ is the amount of compound that causes death of 50% (half) of a group of test animals. It is usually expressed as amount of chemicals administered (e.g., milligrams) per kilogram weight of the test animal. Note that the use of the LD₅₀ data presents some drawbacks when used for QSAR modeling. This is due to the fact that the available data are highly variable and are from different laboratories that do not use the same experimental protocol.

Various QSAR models to predict the acute oral toxicity in rodents were established. Organophosphorus pesticides and herbicides have received particular attention. For example, Enslin (1978) developed regression models using two sets of large data. The R² value for the entire test was 0.33, which means that these models are characterized by low power external prediction. For their part, Adamson et al. (1984) attempted to establish a QSAR model linking LD₅₀ rat oral for 129 herbicide (trifluoromethyl) benzimidazoles and their chemical structure using a multiple regression analysis and the structural descriptors. The authors suggested that reliable forecasts of high precision are not feasible. Nendza (1991) established two models of prediction of the LD₅₀ (mmol/kg) of 12 phenylurea herbicides in rats orally. Hydrophobic and electronic parameters were used and the results obtained are interesting with R values of 0.94 and 0.81.

A set of forty four amide herbicides was used by Zakarya et al. (1996) for establishing two QSAR models for prediction of the LD₅₀. The performance comparison of the ANN model with a three-layer perceptron and a regression model showed that the first model was more efficient. Zakarya et al. (1997) studied structure–toxicity relationships for 120 diverse insecticidal 1, 1, 1-trichloro-2, 2-bis (4-chlorophenyl) ethane-type (DDT-type) molecules using a neural network. Based on the results of the training set, neural networks were found superior to the regression analysis. Eldred and Jurs (1999) proposed two models for predicting the oral LD₅₀ in male rats. The first linear regression-based model focused on a training set of 49 organophosphorus pesticides and a test set consisting of 5 pesticides. The second model was obtained by use of artificial neural networks for 44 and 5 pesticides in learning and test sets, respectively. Twenty nine descriptors were selected from a set of 212 descriptors. The ANN model with seven neurons in the input layer gave RMS values of 0.22 and 0.25 for the sets of learning and testing respectively. In another research work, the acute oral toxicity of 50 amide herbicides in rats (LD₅₀ in mmol/kg) was used to establish a QSAR (Gough and Hall 1999). The model was used to predict the toxicity of new amides, not used throughout the training. The mean absolute error for the test set was equal to 0.27.

QSAR models for prediction of the rat LD₅₀ orally of 67 organophosphorus pesticides (47 in the training set and 20 in the test set) were obtained by Zahouily et al. (2002). Acute toxicity data were converted into mol/kg. Three descriptors were used as inputs of a three layer perceptron trained by backpropagation

algorithm. The optimal neural model has a 3/5/1 structure with $R^2 = 0.93$ and $q^2 = 0.65$. Quantitative structure-toxicity relationship (QSTR) models were derived for estimating the acute oral toxicity of 51 organophosphorus pesticides to male and female rats using regression by partial least squares (PLS) and artificial neural networks (ANN) (Devillers 2004). The first model was able to explain 64% of the variability of the dependent variable. The second nonlinear model obtained with a multilayer perceptron (8/4/1) provided mean square error of 0.29 and 0.26 for the training and test sets respectively. These values were much higher than those of PLS regression. Furthermore, this study has highlighted the importance of the molar refractivity and lipophilicity.

Moreover, Garcia-Domenech and his colleagues (2007) proposed two models for predicting acute toxicity in rats of 62 organophosphorus pesticides. The LD_{50} were expressed in mmol/kg and then log transformed prior to their use. The prediction of the LD_{50} of a test set of 23 pesticides from the second model gave a determination coefficient $R^2 = 0.73$. This work has identified the useful descriptors for developing new QSAR models. Five QSAR models for acute oral toxicity in rats were developed by Zhu et al. (2009a). These models were built using large datasets (253 rats and 235 mice), several methods for statistical modeling and several sets of descriptors. Although the initiative is interesting, these models are not useful in practice owing to the complexity of the development process. Zhu et al. (2009b) also proposed a new modeling approach for predicting acute toxicity. In this study, the authors used chemical descriptors (calculated by DRAGON) and a set of the ZEBET database. The accuracy of the prediction of the LD_{50} of the resulting models exceeds the TOPKAT models applied to the same set of external test.

Recently, Can et al. (2013) have established models for prediction of acute oral toxicity (LD_{50}) of 27 phenyl sulfonylurea herbicides in rats. A method of multilinear regression with four descriptors were selected in this study. The best model gave a value of 0.93 for the coefficient R^2 . The model was validated by internal and external testing. The authors conclude that the test results indicate that the model obtained can be used with confidence to predict the toxicity of molecules phenyl urea herbicides. Hamadache et al. (2014) used multiple linear regression (MLR) and artificial neural network (ANN) to predict acute oral toxicity of a diverse set of 62 herbicides on rats. Both QSAR models obtained using the relevant descriptors showed good predictability. The comparison of results obtained using the ANN model with those of the MLR model revealed the superiority of the ANN model. The statistical parameters for the prediction of acute oral toxicity for MLR and ANN were $R^2 = 0.855$, $RMSE = 0.270$; and $R^2 = 0.960$, $RMSE = 0.118$, respectively. The comparison of the validation results with those of other studies have shown the superiority of the model developed in this work. In a second study for the prediction of acute oral toxicity of 77 herbicides to rats, a QSAR model using an artificial neural network (ANN) was developed (Hamadache et al. 2016a). The internal and external validations of the model showed high values of Q^2 and r_m^2 in the range 0.782–0.997 for training and testing. In addition, the major contribution of the work was to develop an equation based on artificial neural network to predict

the toxicity of 13 other herbicides. The mathematical equation yielded very significant results, which led to an R^2 value of 0.959. The agreement between the calculated and experimental values of acute toxicity confirmed the equation capacity based on ANN to predict the toxicity of herbicides that have not been tested, as well as that of new herbicides.

Very recently, a study on the establishment of a QSAR prediction of acute toxicity of 329 pesticides on rats was undertaken by Hamadache et al. (2016b). The QSAR model based on 17 molecular descriptors is characterized by a good domain of applicability. The best results were obtained with an Artificial Neural Network model with an architecture 17/9/1 established with the Quasi Newton back propagation (BFGS) algorithm. The accuracy of the prediction for the entire external validation was estimated by the Q_{ext}^2 and the mean square error (RMS) being equal to 0.948 and 0.201, respectively. 98.6% Compounds of external validation group were correctly predicted and the current model proved to be superior to previously published models. Consequently, the model developed in that study provides excellent predictions and can be used to predict acute oral toxicity of pesticides, especially for those which have not been tested, as well as new pesticides.

4.3.2 QSARs for Aquatic Toxicity

The acute toxicity to the aquatic environment is determined using a lethal concentration (LC₅₀ 96 h) on fish, median effective concentration (EC₅₀ 48 h) on crustaceans and a median effective concentration (EC₅₀ 72 or 96 h) on alga. The review of the literature shows that the QSAR models dedicated to the aquatic toxicity have been developed mainly for *Daphnia magna* and fathead minnow (*Pimephales promelas*). We note that *Daphnia magna* is classified as the preferred organism for short-term aquatic toxicity testing as suggested in Annex XVII of REACH.

Numbers of QSAR models were developed to predict aquatic toxicity. Some have been developed from a set of homogeneous data, while others from a heterogeneous set. For our part, we will focus on models developed for predicting the aquatic toxicity of pesticides irrespective of the nature of the data sets. An example of these QSAR models can be found in (Agatonovic-Kustrin et al. 2014), where the prediction of the aquatic toxicity of pesticides in terms of lethal dose (LD₅₀) for fish was done using an artificial neural network to a set of 230 pesticides including fungicides, herbicides and insecticides. Thirteen molecular descriptors related to lipophilicity, the hydrogen bond and polarity were selected on 62 calculated descriptors. The authors concluded that this model has predictive power knowing that the value of the predictive coefficient q^2 for the final model was 0.748. Moreover, the experimental values of LC₅₀ to fish of 150 pesticide metabolites (retrieved from the PPDB database: <http://sitem.herts.ac.uk/aeru/ppdb/en/atoz.htm>) were used to develop QSAR model for prediction of acute toxicity using the

software ECOSAR US EPA (Burden et al. 2016). The results show a significant correlation between the values of the predicted and experimental LC_{50} . However, a few outliers were reported. Also, the authors suggest further refining the approach to improve the prediction model and allow future integration in the guidelines and regulatory practices.

Basant et al. (2015a) have established two Quantitative Structure-Toxicity Relationship (QSTR) nonlinear models to predict the toxicity of pesticides for many aquatic species in accordance with the OECD guidelines. A set of six descriptors was used. Model validation was performed using several statistical coefficients for test sets. The two established models applied to data on the toxicity of aquatic species gave R^2 values > 0.92 and 0.97 , respectively. The results suggest the relevance of the QSTR models developed to reliably predict the aquatic toxicity of chemical substances and that they can be used for regulatory purposes. The development of QSAR models to predict the acute toxicity of organothiophosphate pesticides on fish was the objective of the study by Zvinavashe et al. (2009). A set of data on acute toxicity of 15 organothiophosphates to *Daphnia magna* and 3 descriptors were used. In addition, it was examined whether the toxicity data for invertebrate *Daphnia magna* could be used to build a QSAR model to predict toxicity to fish. Appropriate QSAR models (with $0.80 < R^2 < 0.82$) were developed to predict the acute toxicity of organothiophosphates to fish (*Cyprinus carpio*) and to invertebrate (*Daphnia magna*). Internal and external validation was performed on QSAR models and a scope was defined.

A dataset of 125 aromatic pesticides with aquatic toxicity towards trout was used to develop a QSAR model (Slavov et al. 2008). In addition to the standard 2D-QSAR analysis, a comparative molecular field analysis (CoMFA) was also carried out for comparison purposes. The CoMFA analysis contributed to the recognition of steric interactions which play an important role in aquatic toxicity. The QSAR approach initiated by Mazzatorta et al. (2005) was applied for the prediction of acute aquatic toxicity of a set of pesticides. Various linear regression techniques and nonlinear were used to obtain QSAR models. The final model, developed by a counter propagation neural network coupled with genetic algorithms produced good results for the entire test set with $R^2 = 0.79$. Toxicities ($EC_{50} - 24$ h) of 18 biphenyls substituted for *Daphnia magna* were used to develop three linear one descriptor QSAR models (Wang et al. 2004). A good correlation between the predicted and experimental values was noted. However, it was found that the model obtained with quantum chemistry parameter had good predictive ability. Another example of a QSAR model was reported by Devillers (2001). This model was developed using a feed-forward neural network with three layers formed by the back-propagation algorithm; it was used for the prediction of acute toxicity of pesticides against *Lepomis macrochirus* (freshwater fish). However, the authors stress that the model may tend to overfitting for all training data if external validation strategies are not implemented.

Several QSAR models established for homogeneous series of pesticides were used for the prediction of acute toxicity on *Daphnia magna*. For example, Vighi et al. (1991) used the acute toxicity ($EC_{50} - 24$ h) of 22 organophosphorus pesticides to establish a QSAR model based on multi-linear regression. A good correlation was obtained with a coefficient of determination $R^2 = 0.90$. The authors noted that the major contribution to the toxicity for daphnia is the lipophilic nature. Another QSAR study for the prediction of aquatic toxicity ($EC_{50} - 24$ h) involving a set of 20 organophosphate pesticides and WHIM descriptors was conducted by Todeschini et al. (1996). The model based on multi-linear regression provided an excellent predictive power with a coefficient $R^2 = 0.92$. Recently, new QSAR models for predicting the aquatic toxicity ($EC_{50} - 48$ h) of 97 triazole compounds and benzo-triazoles against *Daphnia magna* were established by Cassani et al. (2013). These models, developed by using the multilinear regression were validated in accordance on OECD principles. They are characterized by a strong external predictability ($Q_{ext}^2 = 0.69-0.83$) and a wide applicability domain.

4.3.3 QSARs for Effects on Terrestrial Organisms (Birds, Invertebrates, Plants)

A number of QSAR models have been developed to predict the pesticide toxicity with respect to terrestrial organisms like birds, invertebrates, microorganisms and plants. In this context, we cite some models reported in the literature.

Avian toxicity in four species (Mallard duck, Ring-necked pheasant, Japanese quail, House sparrow) of a set of pesticides was modeled with 9 descriptors and artificial neural networks (Basant et al. 2015b). Three QSAR models (SDT: single decision tree, DTF: decision tree forest, and DTB: decision tree boost) were built according to the OECD guidelines. The second and third models with coefficients $R^2 = 0.945$ and 0.966 respectively gave better prediction results than the first model. The authors emphasize the relevance of QSAR models developed and suggest that they may be useful tools in screening for new pesticides for regulatory purposes. In another study, the QSAR analyses for fungicidal activities of thiazoline derivatives against rice blast (*Magnaporthe grisea*) were developed with physico-chemical descriptors using multiple linear regression (MLR) and neural network (NN) (Song et al. 2008). Three sets of thiazoline derivatives with different substitution patterns were used. The models developed according to the OECD principles were subjected to internal and external validation showing good results. For example, a sample consisting of 82 compounds in the training set was accredited with a standard error equal to 0.097 (ANN) and 0.139 (MLR), whereas the standard error values for the test set were 0.122 (ANN) and 0.162 (MLR).

Devillers et al. (2002) have developed a feed-forward neural model for the prediction of acute toxicity of 100 pesticides to *Apis mellifera* (European bee) by using the physicochemical properties as descriptors. The root mean square residual (RMSR) values for the training and testing sets were 0.430 and 0.386, respectively.

The model developed according to the principles 1 and 4 of the OECD underwent internal and external validation tests. Neural models of quantitative structure-toxicity relationship were established for predicting the qualitative and quantitative toxicity of pesticides in honey bee (*Apis mellifera*) using experimental values of the toxicity of 237 pesticides (Singh et al. 2014). The predictive power of the models was tested by internal and external validation with different statistical metrics. One of the models gave a correlation value of R^2 of 0.841 and a mean squared error (MSE) of 0.22. The authors suggest that the two built models can be useful tools for predicting the qualitative and quantitative toxicities of new pesticides for regulatory purposes.

5 Conclusions

Based on the review of the literature on QSAR models dedicated to predicting the toxicity of pesticides in the context of risk assessment, two types of models are currently available: those generated by commercial software and those published in the literature. Since a lot of literature has been devoted to the application of commercial software in predicting toxicity, our literature review was devoted essentially to the QSAR models published in the literature.

Critical evaluation of the QSAR models for predicting the toxicity of pesticides that are reported so far in the literature allowed us to draw the following conclusions. First, to achieve the objectives established under the REACH inherent to toxicity of chemical compounds, efforts are necessary and imperative to develop very effective in silico prediction methods. Among these methods, we can cite quantitative structure-activity relationships, which are essential. Then, thanks to the remarks made by some authors (Stouch et al. 2003; Johnson 2008) on the feasibility and reliability of the use of QSAR approaches in toxicity studies: QSAR models should be established strictly under the principles established by the OECD. Finally, QSAR methods based on neural networks have also provided promising results in predicting the toxicity of compounds with respect to certain species. These models can be faster and less expensive, and are alternatives to toxicity testing involving animal experiments.

In this review, a number of QSAR models for prediction of certain toxicological properties of pesticides are summarized. Compared with the number of models available in the literature for predicting the toxicity of chemicals, those dedicated to pesticides remain insignificant. Furthermore, the study revealed considerable imbalances in the availability of models compared to the toxicological endpoint of regulatory significance studied. In one hand, there is abundant literature for the prediction of acute toxicity and aquatic toxicity and on the other hand, there is a few or no QSAR models devoted to the prediction of other toxicity properties.

In general, most QSAR models for the prediction of acute and aquatic toxicities available in the literature were drawn from limited data sets of pesticides that have similar chemical structures, such as organophosphates. However, some models

derived from heterogeneous data sets have also been reported in the literature (Benfenati et al. 2007; Zhu et al. 2009a; Hamadache et al. 2016b; Agatonovic-Kustrin et al. 2014; Burden et al. 2016). In addition, the number of statistical parameters used for validation is limited, especially in the case of older old works. Moreover, these models are acceptable with regard to some of the OECD guidelines for validation of QSAR, especially the first and the fourth principles. Furthermore, in the regulatory assessment of pesticides, most of the established models are far from meeting the requirements of other OECD validation principles. Therefore, despite the fact that the QSAR models developed in some very recent studies are promising in that they showed very interesting predictive capabilities, efforts are needed to explore their applicability and implement them in a useful form in the practice. Concerning the parameters having a great contribution to the toxicity of pesticides, hydrophobicity, steric effects and electronic effects, can be mentioned. Moreover, they can be useful in improving the understanding of the mechanisms involved in the toxicity of the substances studied.

References

- Adamson, G. W., Bawden, D., & Siggers, D. T. (1984). Quantitative structure–activity relationship studies of acute toxicity (LD50) in a large series of herbicidal benzimidazoles. *Pesticide Science*, 15, 31–39.
- Agatonovic-Kustrin, S., Morton, D. W., & Razic, S. (2014). In silico modelling of pesticide aquatic toxicity. *Combinatorial Chemistry and High Throughput Screen*, 17(9), 808–818.
- Agrawal, A., Pandey, R. S., & Sharma, B. (2010). Water pollution with special reference to pesticide contamination in India. *Journal of Water Resource Protection*, 2, 432–448.
- Ahouangninou, C., Thibaud, M., Edoth, P., et al. (2012). Characterization of health and environmental risks of pesticide use in market-gardening in the rural city of Tori-Bossito in Benin, West Africa. *Journal of Environmental Protection*, 3, 241–248.
- Ahsan, H., Karim, N., Sanwer Ali, S., et al. (2013). Impact of pesticides contamination on nutritional values of marine fishery from Karachi Coast of Arabian Sea. *Food and Nutrition Sciences*, 4, 924–932.
- AIRPARIF. (2007). Surveillance de la qualité de l'air en Ile de France. Evaluation des concentrations en pesticides dans l'air ambiant francilien: Campagne exploratoire.
- Akan, J. C., Mohammed, Z., Jafiya, L., et al. (2013). Organochlorine pesticide residues in fish samples from Alau Dam, Borno State, North Eastern Nigeria. *Journal of Environmental and Analytical Toxicology*, 3, 171. doi:10.4172/2161-0525.1000171.
- Ali, U., Jabir Hussain, S., Riffat Naseem, M., et al. (2014). Organochlorine pesticides (OCPs) in South Asian region: A review. *Science of the Total Environment*, 476–477, 705–717.
- Andersen, H. R., Ida, M. S., Grandjean, P., et al. (2008). Impaired reproductive development in sons of women occupationally exposed to pesticides during pregnancy. *Environmental Health Perspectives*, 116(4), 566–572.
- Arias, L. A., Bojaca, C., Ahumada, A. D., et al. (2014). Monitoring of pesticide residues in tomato marketed in Bogota, Colombia. *Food Control*, 35, 213–217.
- Arzi, A., Hemmati, A. A., & Nazari Khorasgani, Z. (2011). Determination and comparison of the organochlorine pesticide residue levels among benni fish of Shadegan, mahshahr and susangerd cities, Khozestan province in Iran. *Jundishapur Journal of Natural Pharmaceutical Products*, 6(1), 24–31.

- Asman, W. A. H., Jorgensen, A., Bossi, R., et al. (2005). Wet deposition of pesticides and nitrophenols at two sites in Denmark: Measurements and contributions from regional sources. *Chemosphere*, *59*, 1023–1031.
- Baldi, I., Lebailly, P., Rondeau, V., et al. (2012). Levels and determinants of pesticide exposure in operators involved in treatment of vineyards: Results of the PESTEXPO Study. *Journal of Exposure Science and Environmental Epidemiology*, *22*(6), 593–600.
- Band, P. R., Abanto, Z., Bert, J., et al. (2011). Prostate cancer risk and exposure to pesticides in British Columbia farmers. *Prostate*, *71*(2), 168–183.
- Barganska, Z., Slebioda, M., & Namiesnik, J. (2013). Pesticide residues levels in honey from apiaries located of Northern Poland. *Food Control*, *31*, 196–201.
- Barjhoux, I. (2011). Thèse de Doctorat: Étude de la biodisponibilité et de la toxicité de polluants chimiques à risque dans les sédiments aquatiques vis-à-vis des premiers stades de développement d'un poisson modèle, *Oryzias latipes*. Université Bordeaux 1, France.
- Basant, N., Gupta, S., & Singh, K. P. (2015a). Predicting toxicities of diverse chemical pesticides in multiple avian species using tree-based QSAR approaches for regulatory purposes. *Journal of Chemical Information and Modeling*, *55*(7), 1337–1348. doi:10.1021/acs.jcim.5b00139.
- Basant, N., Gupta, S., & Singh, K. P. (2015b). Predicting aquatic toxicities of chemical pesticides in multiple test species using nonlinear QSTR modeling approaches. *Chemosphere*, *139*, 246–255. doi:10.1016/j.chemosphere.2015.06.063.
- Bedos, C., Cellier, P., Calvet, R., et al. (2002). Mass transfer of pesticides into the atmosphere by volatilization from soils and plants: Overview. *Agronomie*, *22*, 21–33.
- Bempah, C. K., Asomaning, J., & Boateng, J. (2012). Market basket survey for some pesticides residues in fruits and vegetables from Ghana. *J microbiol biotechnol food sci*, *2*(3), 850–871.
- Benfenati, E., Craciun, M., & Neagu, D. (2007). The use of the DEMETRA models. In E. Benfenati (Ed.), *Quantitative structure-activity relationship (QSAR) for pesticide regulatory purposes* (pp. 303–313). Amsterdam: Elsevier.
- Borras, E., Sanchez, P., Munoz, A., et al. (2011). Development of a gas chromatography-mass spectrometry method for the determination of pesticides in gaseous and particulate phases in the atmosphere. *Analytica Chimica Acta*, *699*, 57–65.
- Bottoni, P. P., Grenni, L., Lucentini, A., et al. (2013). Terbutylazine and other triazines in Italian water resources. *Microchemical Journal*, *107*, 136–142.
- Bretveld, R., Kik, S., Hooiveld, M., et al. (2008). Time-to pregnancy among male greenhouse workers. *Occupational and Environmental Medicine*, *65*, 185–190.
- Briand, O. (2003). Influence des facteurs environnementaux et des pratiques agricoles sur les variations spatio-temporelles des niveaux de contamination de l'atmosphère par les pesticides (Doctoral dissertation) Rennes 1 France.
- Brophy, J. T., Keith, M. M., Gorey, K. M., et al. (2002). Occupational histories of cancer patients in a canadian cancer treatment center and the generated hypothesis regarding breast cancer and farming. *International Journal of Occupational Medicine and Environmental Health*, *8*, 346–353.
- Burden, N., Maynard, S. K., Weltje, L., et al. (2016). The utility of QSARs in predicting acute fish toxicity of pesticide metabolites: A retrospective validation approach. *Regulatory Toxicology and Pharmacology pii*, *S0273–2300*(16), 30146–5. doi:10.1016/j.yrtph.2016.05.032.
- Cabidoche, Y. M., & Lesueur-Jannoyer, M. (2012). Contamination of harvested organs in root crops grown on chlordecone-polluted soils. *Pedosphere*, *22*(4), 562–571.
- Cachot, J. (2014). Assessment of pollution in the Bizerte lagoon (Tunisia) by the combined use of chemical and biochemical markers in mussels, *Mytilus gallo-provincialis*. *Marine Pollution Bulletin*, *84*, 379–390.
- Calatayud-Vernich, P., Calatayud, F., Simo, E., et al. (2016). Influence of pesticide use in fruit orchards during blooming on honeybee mortality in 4 experimental apiaries. *Science of the Total Environment*, *41*, 33–41.
- Can, A., Yildiz, I., & Guvendik, G. (2013). The determination of toxicities of sulphonylurea and phenylurea herbicides with quantitative structure-toxicity relationship (QSTR) studies. *Environmental Toxicology and Pharmacology*, *35*(3), 369–379.

- Cassani, S., Kovarich, S., Papa, E., et al. (2013). Daphnia and fish toxicity of (benzo) triazoles: Validated QSAR models, and interspecies quantitative activity–activity modelling. *Journal of Hazardous Materials*, 258(259), 50–60.
- Cassotti, M., Consonni, V., Mauri, A., et al. (2014). Validation and extension of a similarity-based approach for prediction of acute aquatic toxicity towards *Daphnia magna*. *SAR and QSAR in Environmental Research*, 25, 1013–1036.
- Chai, L. K., Wong, M. H., & Bruun Hansen, H. C. (2013). Degradation of chlorpyrifos in humid tropical soils. *Journal of Environmental Management*, 125, 28–32.
- Chevrier, C., Limon, G., Monfort, C., et al. (2011). Urinary biomarkers of prenatal atrazine exposure and adverse birth outcomes in the PELAGIE Birth Cohort. *Environmental Health Perspectives*, 119, 1034–1041.
- Chrisman, J. R., Koifman, S., de Novaes, Sarcinelli P., et al. (2009). Pesticide sales and adult male cancer mortality in Brazil. *International Journal of Hygiene and Environmental Health*, 212, 310–321.
- Christiansen, S., Boberg, J., Nelleman, C., et al. (2010). A cocktail of endocrine disrupting pesticides affects sexual differentiation in rats. *Reproductive Toxicology*, 30(2), 229–229.
- Coat, S., Monti, D., Legendre, P., et al. (2011). Organochlorine pollution in tropical rivers (Guadeloupe): Role of ecological factors in food web bioaccumulation. *Environmental Pollution*, 159(6), 1692–1701. doi:10.1016/j.envpol.2011.02.036.
- Coscolla, C., Colin, P., Yahyaoui, A., et al. (2010). Occurrence of currently used pesticides in ambient air of center region (France). *Atmospheric Environment*, 44, 3915–3925.
- Coscolla, C., Hart, E., Pastor, A., et al. (2013). LC-MS characterization of contemporary pesticides in PM10 of Valencia Region, Spain. *Atmospheric Environment*, 77, 394–403.
- Costello, S., Cockburn, M., Bronstein, J., et al. (2009). Parkinson's disease and residential exposure to Maneb and Paraquat From agricultural applications in the Central Valley of California. *American Journal of Epidemiology*, 169, 919–926.
- Coupe, R. H., & Blomquist, J. D. (2004). Water-soluble pesticides in finished water of community water supplies. *Journal of AWWA*, 96, 56–68.
- Cruzeiro, C., Rocha, E., Pardal, M. A., et al. (2016). Environmental assessment of pesticides in the Mondego River Estuary (Portugal). *Marine Pollution Bulletin*, 103, 240–246.
- Dabrowski, J. M., Shadung, J. M., & Wepener, V. (2014). Prioritizing agricultural pesticides used in South Africa based on their environmental mobility and potential human health effects. *Environment International*, 62, 31–40.
- Dalvie, M. A., & London, L. (2009). Risk assessment of pesticide residues in South African raw wheat. *Crop Protection*, 28, 864–869.
- Davodi, M., Esmaili-Sari, A., & Bahramifarr, N. (2011). Concentration of polychlorinated biphenyls and organochlorine pesticides in some edible fish species from the Shadegan Marshes (Iran). *Ecotoxicology and Environmental Safety*, 74, 294–300.
- De Gerónimo, E., Aparicio, V. C., Bárbaro, S., et al. (2014). Presence of pesticides in surface water from four sub-basins in Argentina. *Chemosphere*, 107, 423–431.
- Dearden, J. C. (2002). Prediction of environmental toxicity and fate using quantitative structure-activity relationships (QSARs). *Journal of the Brazilian Chemical Society*, 13(6), 754–762.
- Dearden, J. C., & Rowe, P. H. (2015). Use of artificial neural networks in the QSAR prediction of physicochemical properties and toxicities for REACH legislation (chapter 5). In H. Cartwright (Ed.), *Artificial neural networks, Methods in molecular biology* (Vol. 1260). doi:10.1007/978-1-4939-2239-0_5.
- Devillers, J. (2001). A general QSAR model for predicting the acute toxicity of pesticides to *Lepomis macrochirus*. *SAR and QSAR in Environmental Research*, 11, 397–417.
- Devillers, J. (2004). Prediction of mammalian toxicity of organophosphorus pesticides from QSTR modeling. *SAR and QSAR in Environmental Research*, 15(5), 501–510.
- Devillers, J., & Devillers, H. (2009). Prediction of acute mammalian toxicity from QSARs and interspecies correlations. *SAR and QSAR in Environmental Research*, 20(5–6), 467–500.

- Devillers, J., Pham-Delègue, M. H., Decourtye, A., et al. (2002). Structure-toxicity modeling of pesticides to honey bees. *SAR and QSAR in Environmental Research*, 13(7–8), 641–648.
- Dirinck, E. L., Dirtu, A. C., Govindan, M., et al. (2014). Exposure to persistent organic pollutants: Relationship with abnormal glucose metabolism and visceral adiposity. *Diabetes Care*, 37, 1951–1958.
- Eldred, D. V., & Jurs, P. C. (1999). Prediction of acute mammalian toxicity of organophosphorus pesticide compounds from molecular structure. *SAR and QSAR in Environmental Research*, 10(2), 75–99.
- Enslin, K. (1978). A toxicity estimation model. *Journal of Environmental Pathology and Toxicology*, 2(1), 115–121.
- EU. (2006). *Official Journal of the European Union L 396, 49*, Regulation (EC) N 1907/2006 Article 13 General requirements or generation of information on intrinsic properties of substances.
- Evangelou, E., Ntritsos, G., Chondrogiorgi, M., et al. (2016). Exposure to pesticides and diabetes: A systematic review and meta-analysis. *Environment International*, 91, 60–68.
- Farajzadeh, M. A., Khoshmaram, L., & Alizadeh Nabil, A. A. (2014). Determination of pyrethroid pesticides residues in vegetable oils using liquid–liquid extraction and dispersive liquid–liquid microextraction followed by gas chromatography–flame ionization detection. *Journal of Food Composition and Analysis*, 34(2), 128–135.
- Feng, J., Tang, H., Chen, D., et al. (2015). Monitoring and risk assessment of pesticide residues in tea samples from China. *Human and Ecological Risk Assessment*, 21, 169–183.
- Fianko, J. R., Donkor, A., Lowor, S. T., et al. (2011). Health risk associated with pesticide contamination of fish from the Densu River Basin in Ghana. *Journal of Environmental Protection*, 2, 115–123.
- Fischer, W. J., Schilter, B., Tritscher, A.M., et al. (2011). Contaminants of milk and dairy products: Contamination resulting from farm and dairy practices. *Encyclopedia of Dairy Sciences* (2nd ed., pp. 887–897).
- Floch, C., Chevremont, A. C., Joanico, K., et al. (2011). Indicators of pesticide contamination: Soil enzyme compared to functional diversity of bacterial communities via Biolog Ecoplates. *European Journal of Soil Science*, 47, 256–263.
- Freire, C., & Koifman, S. (2012). Pesticide exposure and Parkinson's disease: Epidemiological evidence of association. *Neurotoxicology*, 33(5), 947–971.
- Furlong, M., Tanner, C. M., Goldman, S. M., et al. (2015). Protective glove use and hygiene habits modify the associations of specific pesticides with Parkinson's disease. *Environment International*, 75, 144–150.
- Galiulin, R. V., Bashkin, V. N., Galiulina, R. A., et al. (2002). Behavior of persistent organic pollutants in the airplant-soil system. *Water, Air, and Soil Pollution*, 37, 179–191.
- Gao, J., Zhou, H., Pan, G., et al. (2013). Factors influencing the persistence of organochlorine pesticides in surface soil from the region around the Hongze Lake, China. *Science of the Total Environment*, 443, 7–13.
- Garcia-Domenech, R., Alarcon-Elbal, P., Bolas, G., et al. (2007). Prediction of acute toxicity of organophosphorus pesticides using topological indices. *SAR and QSAR in Environmental Research*, 18(7), 745–755.
- Ge, J., Woodward, L. A., Li, Q. X., et al. (2013). Composition, distribution and risk assessment of organochlorine pesticides in soils from the Midway Atoll, North Pacific Ocean. *Science of the Total Environment*, 452, 421–426.
- Golbamaki, A., Cassano, A., Lombardo, A., et al. (2014). Comparison of in silico models for prediction of *Daphnia magna* acute toxicity. *SAR and QSAR in Environmental Research*, 25, 673–694.
- Gough, J. D., & Hall, L. H. (1999). Modeling the toxicity of amide herbicides using the electrotopological state. *Environmental Toxicology and Chemistry*, 18, 1069–1075.

- Grote, K., Niemann, L., Selzsam, B., et al. (2008). Epoxiconazole causes changes in testicular histology and sperm production in the Japanese quail (*Coturnix coturnix japonica*). *Environmental Toxicology and Chemistry*, 27, 2368–2374.
- Grung, M., Lin, Y., Zhang, H., et al. (2015). Pesticide levels and environmental risk in aquatic environments in China—A review. *Environment International*, 81, 87–97.
- Gryniewicz, M., Polkowska, Z., Gorecki, T., et al. (2001). Pesticides in precipitation in the Gdansk region (Poland). *Chemosphere*, 43(3), 303–312.
- Gunier, R. B., Ward, M. H., Airola, M., et al. (2011). Determinants of agricultural pesticide concentrations in carpet dust. *Environmental Health Perspectives*, 119, 970–976.
- Haddaoui, I., Olfa, M., Borhane, M., et al. (2016). Occurrence and distribution of PAHs, PCBs, and chlorinated pesticides in Tunisian soil irrigated with treated wastewater. *Chemosphere*, 146, 195–205.
- Hamadache, M., Benkortbi, O., Hanini, S., et al. (2016a). A quantitative structure activity relationship for acute oral toxicity of pesticides on rats: Validation, domain of application and prediction. *Journal of Hazardous Materials*, 303, 28–40.
- Hamadache, M., Hanini, S., Benkortbi, O., et al. (2016b). Artificial neural network-based equation to predict the toxicity of herbicides on rats. *Chemometrics and Intelligent Laboratory Systems*, 154, 7–15.
- Hamadache, M., Khaouane, L., Benkortbi, O., et al. (2014). Prediction of acute herbicide toxicity in rats from quantitative structure-activity relationship modeling. *Environmental Engineering Science*, 31(5), 243–252.
- Hart, E., Coscolla, C., Pastor, A., et al. (2012). GC-MS characterization of contemporary pesticides in PM10 of Valencia region, Spain. *Atmospheric Environment*, 62, 118–129.
- Hayden, K. M., Norton, M. C., Darcey, D., et al. (2010). Occupational exposure to pesticides increases the risk of incident AD: The Cache County study. *Neurology*, 74(19), 1524–1530.
- Hina, A., Nasim, K., Syed S. A., Munshi, A. B., Shaukat, S. (2013). Impact of pesticides contamination on nutritional values of marine fishery from Karachi Coast of Arabian Sea. *Food and Nutrition Sciences*, 4, 924–932.
- Hogarth, J. N., Seike, N., Kobara, Y., et al. (2013). Seasonal variation of atmospheric polychlorinated biphenyls and polychlorinated naphthalenes in Japan. *Atmospheric Environment*, 80, 275–280.
- Hoppin, J. A., Umbach, D. M., London, S. J., et al. (2002). Chemical predictors of wheeze among farmer pesticide applicators in the agricultural health study. *American Journal of Respiratory and Critical Care Medicine*, 165(5), 683–689.
- Hoppin, J. A., Valcin, M., Henneberger, P. K., et al. (2007). Pesticide use and chronic bronchitis among farmers in the agricultural health study. *American Journal of Industrial Medicine*, 50(12), 969–979.
- Jaacks, L. M., & Staimez, L. R. (2015). Association of persistent organic pollutants and non-persistent pesticides with diabetes and diabetes-related health outcomes in Asia: A systematic review. *Environment International*, 76, 57–70.
- Jiang, Y. F., Wang, X. T., Wu, M. H., et al. (2011). Contamination, source identification, and risk assessment of polycyclic aromatic hydrocarbons in agricultural soil of Shanghai. *China Environ Monit Assess*, 183, 139–150.
- Johnson, S. R. (2008). The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *Journal of Chemical Information and Modeling*, 48(1), 25–26.
- Juan-Borras, M., Domenech, E., & Escriche, I. (2016). Mixture-risk-assessment of pesticide residues in retail polyfloral honey. *Food Control*, 67, 127–134.
- Kjaerstad, M. B., Taxvig, C., NELLEMAN, C., et al. (2010). Endocrine disrupting effects in vitro of conazoles antifungals used as pesticides and pharmaceuticals. *Reproductive Toxicology*, 30, 573–582.
- Koureas, M., Tsakalof, A., Tsatsakis, A., et al. (2012). Systematic review of biomonitoring studies to determine the association between exposure to organo-phosphorus and pyrethroid insecticides and human health outcomes. *Toxicology Letters*, 210, 155–168.

- Lagunin, A., Zakharov, A., Filimonov, D., et al. (2011). QSAR modelling of rat acute toxicity on the basis of PASS prediction. *Molecular Informatics*, 30, 241–250. doi:10.1002/minf.201000151.
- Lebov, J. F., Engel, L. S., Richardson, D., et al. (2016). Pesticide use and risk of end-stage renal disease among licensed pesticide applicators in the agricultural health study. *Occupational and Environmental Medicine*, 73, 3–12.
- Lee, I., Eriksson, P., Fredriksson, A., et al. (2015). Developmental neurotoxic effects of two pesticides: Behavior and biomolecular studies on chlorpyrifos and carbaryl. *Toxicology and Applied Pharmacology*, 288, 429–438.
- Lerro, C. C., Koutros, S., Andreotti, G., et al. (2015). Organophosphate insecticide use and cancer incidence among spouses of pesticide applicators in the agricultural health study. *Occupational and Environmental Medicine*, 72, 736–744.
- Liu, D., & Min, S. (2012). Rapid analysis of organochlorine and pyrethroid pesticides in tea samples by directly suspended droplet microextraction using a gas chromatography–electron capture detector. *Journal of Chromatography A*, 1235, 166–173.
- Liu, Y., Shen, D., Li, S., et al. (2016a). Residue levels and risk assessment of pesticides in nuts of China. *Chemosphere*, 144, 645–651.
- Liu, W. X., Wang, Y., He, W., et al. (2016b). Aquatic biota as potential biological indicators of the contamination, bioaccumulation and health risks caused by organochlorine pesticides in a large, shallow Chinese lake (Lake Chaohu). *Ecological Indicators*, 60, 335–345.
- Lozowicka, B., Kaczynski, P., Paritova, A. E., et al. (2014). Pesticide residues in grain from Kazakhstan and potential health risks associated with exposure to detected pesticides. *Food and Chemical Toxicology*, 64, 238–248.
- Ma, X. X., & Ran, Y. (2009). The research for organochlorine pesticides in soils of the Pearl River Delta. *Ecological Environmental Sciences*, 18(1), 134–137.
- Mas, S., de Juan, A., Tauler, R., et al. (2010). Application of chemometric methods to environmental analysis of organic pollutants: A review. *Talanta*, 80, 1052–1067.
- Mast, M. A., Campbell, D. H., Ingersoll, G. P., et al. (2003). Atmospheric deposition of nutrients, pesticides, and mercury in Rocky Mountain National Park, Colorado, 2002 US Department of the Interior, U.S. Geological Survey Water-Resources Investigations Report 03-4241.
- Mazzatorra, P., Smiesko, M., Lo Piparo, E., et al. (2005). QSAR model for predicting pesticide aquatic toxicity. *Journal of Chemical Information and Modeling*, 45(6), 1767–1774.
- McKinlay, R., Plant, J. A., Bell, J. N. B., et al. (2008). Endocrine disrupting pesticides: Implications for risk assessment. *Environment International*, 34, 168–183.
- Mills, P. K., & Yang, R. C. (2007). Agricultural exposures and gastric cancer risk in Hispanic farm workers in California. *Environmental Research*, 104, 282–289.
- MOEJ: Ministry of the Environment of Japan. (2015). *Chemicals in the environment* (p 648).
- Montgomery, M. P., Kamel, F., Saldana, T. M., et al. (2008). Incident diabetes and pesticide exposure among licensed pesticide applicators: Agricultural health study 1993–2003. *American Journal of Epidemiology*, 67, 1235–1246.
- Mostafalou, S., & Abdollahi, M. (2013). Pesticides and human chronic diseases: Evidences, mechanisms, and perspectives. *Toxicology and Applied Pharmacology*, 268(2), 157–177. doi:10.1016/j.taap.2013.01.025.
- Moussaoui, Y., Tuduri, L., Kerchich, Y., et al. (2012). Atmospheric concentrations of PCDD/Fs, dl-PCBs and some pesticides in Northern Algeria using passive air sampling. *Chemosphere*, 88, 270–277.
- Multigner, L. (2005). Effets retardés des pesticides sur la santé humaine. *Environnement, Risques et Santé*, 3, 187–194.
- Munaron, D. (2004). *Etude des apports en herbicides et en nutriments par la CHARENTE: modélisation de la dispersion de l'atrazine dans le bassin de MARENNE-LERON*. Thèse de Doctorat: Université Pierre et Marie Curie, Paris VI France.
- Nandi, S., Gupta, P. S., Roy, S. C., et al. (2011). Chlorpyrifos and endosulfan affect buffalo oocyte maturation, fertilization, and embryo development in vitro directly and through cumulus cells. *Environmental Toxicology*, 26(1), 57–67.

- Nendza, M. (1991). Predictive QSAR models estimating ecotoxic hazard of phenylureas: Mammalian toxicity. *Chemosphere*, 22(5–6), 613–623.
- Nougadère, A., Sirof, V., Kadar, A., et al. (2012). Total diet study on pesticide residues in France: Levels in food as consumed and chronic dietary risk to consumers. *Environment International*, 45, 135–150.
- Orton, F., Rosivatz, E., Scholze, M., et al. (2011). Widely used pesticides with previously unknown endocrine activity revealed as in vitro antiandrogens. *Environmental Health Perspectives*, 119(6), 794–800.
- Oukali-Haouchine, O., Barriuso, E., Mayata, Y., et al. (2013). Factors affecting Métribuzine retention in Algerian soils and assessment of the risks of contamination. *Environmental Monitoring and Assessment*, 185, 4107–4115.
- Palma, P., Köck-Schulmeyer, M., Alvarenga, P., et al. (2014). Risk assessment of pesticides detected in surface water of the Alqueva reservoir (Gadiana basin, southern of Portugal). *Science of the Total Environment*, 488–489, 208–219.
- Papadakis, E. N., Tسابoula, A., Kotopoulou, A., et al. (2015). Pesticides in the surface waters of Lake Vistonis Basin, Greece: Occurrence and environmental risk assessment. *Science of the Total Environment*, 536, 793–802.
- Parron, T., Requena, M., Hernandez, A. F., et al. (2011). Association between environmental exposure to pesticides and neurodegenerative diseases. *Toxicology and Applied Pharmacology*, 256(3), 379–385.
- Peshin, R., & Zhang, W. J. (2014). Integrated pest management and pesticide use (Chapter 1). *Integrated pest management* (Vol. 3, pp. 1–46). Heidelberg: Springer.
- Polkowska, Z., Kot, A., Wiergowski, M., et al. (2000). Organic pollutants in precipitation: Determination of pesticides and polycyclic aromatic hydrocarbons in Gdansk, Poland. *Atmospheric Environment*, 34(8), 1233–1245.
- Provost, D., Cantagrel, A., Lebailly, P., et al. (2007). Brain tumours and exposure to pesticides: A case–control study in Southwestern France. *Occupational and Environmental Medicine*, 64(8), 509–514.
- Qu, C., Qi, S., Yang, D., et al. (2015). Risk assessment and influence factors of organochlorine pesticides (OCPs) in agricultural soils of the hill region: A case study from Ningde, southeast China. *Journal of Geochemical Exploration*, 149, 43–51.
- Raepfel, C., Fabritius, M., Nief, M., et al. (2014). Coupling ASE, silylation and SPME-GC/MS for the analysis of current-used pesticides in atmosphere. *Talanta*, 121, 24–29.
- Rasmussen, J. J., Wiberg-Larsen, P., Baattrup-Pedersen, A., et al. (2015). The legacy of pesticide pollution: An overlooked factor in current risk assessments of freshwater systems. *Water Research*, 84, 25–32.
- Rebich, R. A., Coupe, R. H., & Thurma, E. M. (2004). Herbicide concentrations in the Mississippi River Basin, the importance of chloroacetanilide herbicide degradates. *Science of the Total Environment*, 321, 189–199.
- Saeed, T., Sawaya, W. N., Ahmad, N., et al. (2005). Organophosphorus pesticide residues in the total diet of Kuwait. *Arabian Journal of Science and Engineering*, 30(1A), 17–27.
- Saeedi Saravi, S. S., & Dehpour, A. R. (2016). Potential role of organochlorine pesticides in the pathogenesis of neurodevelopmental, neurodegenerative, and neurobehavioral disorders: A review. *Life Sciences*, 145, 255–264.
- Sanagi, M. M., Salleh, S., Ibrahim, W. A. W., et al. (2013). Molecularly imprinted polymer solid-phase extraction for the analysis of organophosphorus pesticides in fruit samples. *Journal of Food Composition and Analysis*, 32, 155–161.
- Sazonovas, A., Japertas, P., Didziapetris, R., et al. (2010). Estimation of reliability of predictions and model applicability domain evaluation in the analysis of acute toxicity (LD50). *SAR and QSAR in Environmental Research*, 21, 127–148.
- Scheyer, A., Graeff, C., Morville, S., et al. (2005). Analysis of some organochlorine pesticides in an urban atmosphere (Strasbourg, east of France). *Chemosphere*, 58(11), 1517–1524.

- Schummer, C., Mothiron, E., Appenzeller, B. M. R., et al. (2010). Temporal variations of concentrations of currently used pesticides in the atmosphere of Strasbourg, France. *Environmental Pollution*, *158*, 576–584.
- Shoiful, A., Fujita, H., Watanabe, I., et al. (2013). Concentrations of organochlorine pesticides (OCPs) residues in foodstuffs collected from traditional markets in Indonesia. *Chemosphere*, *90*, 1742–1750.
- Shojaei Saadi, H., & Abdollahi, M. (2012). Is there a link between human infertilities and exposure to pesticides. *International Journal of Pharmacology*, *8*(8), 708–710.
- Silva, E., Daam, M. A., & Cerejeira, M. J. (2015). Aquatic risk assessment of priority and other river basin specific pesticides in surface waters of Mediterranean river basins. *Chemosphere*, *135*, 394–402.
- Singh, K. P., Gupta, S., Basant, N., et al. (2014). QSTR modeling for qualitative and quantitative toxicity predictions of diverse chemical pesticides in honey bee for regulatory purposes. *Chemical Research in Toxicology*, *27*(9), 1504–1515.
- Skretteberg, L. G., Lyrån, B., Holen, B., et al. (2015). Pesticide residues in food of plant origin from Southeast Asia—A Nordic project. *Food Control*, *51*, 225–235.
- Slavov, S., Gini, G., & Benfenati, E. (2008). QSAR trout toxicity models on aromatic pesticides. *Journal of Environmental Science and Health. Part B: Pesticides, Food Contaminants, and Agricultural Wastes*, *43*(8), 633–637. doi:10.1080/10934520801893725.
- Song, J. S., Moon, T., Nam, K. D., et al. (2008). Quantitative structure-activity relationship (QSAR) studies for fungicidal activities of thiazoline derivatives against rice blast. *Bioorganic & Medicinal Chemistry Letters*, *18*, 2133–2142.
- Steen, R. J., Van der Vaart, J., Hiep, M., et al. (2001). Gross fluxes and estuarine behaviour of pesticides in the Scheldt Estuary (1995–1997). *Environmental Pollution*, *115*(1), 65–79.
- Stouch, T. R., Kenyon, J. R., Johnson, S. R., et al. (2003). In silico ADME/Tox: Why models fail. *Journal of Computer-Aided Molecular Design*, *17*(2–4), 83–92.
- Sullivan, K. M., Manuppello, J. R., & Willett, C. E. (2014). Building on a solid foundation: SAR and QSAR as a fundamental strategy to reduce animal testing. *SAR and QSAR in Environmental Research*, *25*, 357–365.
- Sun, J., Pan, L., Zhan, Y., et al. (2016a). Contamination of phthalate esters, organochlorine pesticides and polybrominated diphenyl ethers in agricultural soils from the Yangtze River Delta of China. *Science of the Total Environment*, *544*, 670–676.
- Sun, H., Qi, Y., Zhang, D., et al. (2016b). Concentrations, distribution, sources and risk assessment of organohalogenated contaminants in soils from Kenya, Eastern Africa. *Environmental Pollution*, *209*, 177–185.
- Tagert, M. L., Massey, J. H., & Shaw, D. (2014). Water quality survey of Mississippi's Upper Pearl River. *Science of the Total Environment*, *481*, 564–573.
- Takazawa, Y., Takasuga, T., Doi, K. et al. (2016). Recent decline of DDTs among several organochlorine pesticides in background air in East Asia. *Environmental Pollution*. doi:10.1016/j.envpol.2016.02.019.
- Todeschini, R., Vighi, M., Provenzani, R., et al. (1996). Modeling and prediction by using WHIM descriptors in QSAR studies: Toxicity of heterogeneous chemicals on *Daphnia magna*. *Chemosphere*, *32*, 1527–1545.
- Tsakiris, I. N., Goumenou, M., Tzatzarakis, M. N., et al. (2015). Risk assessment for children exposed to DDT residues in various milk types from the Greek market. *Food and Chemical Toxicology*, *75*, 156–165.
- Van der Mark, M., Brouwer, M., Kromhout, H., et al. (2012). Is pesticide use related to Parkinson disease? Some clues to heterogeneity in study results. *Environmental Health Perspectives*, *120* (3), 340–347.
- Van Maele-Fabry, G., Hoet, P., & Lison, D. (2013). Parental occupational exposure to pesticides as risk factor for brain tumors in children and young adults: A systematic review and meta-analysis. *Environment International*, *56*, 19–31.

- Van Maele-Fabry, G., Hoet, P., Vilain, F., et al. (2012). Occupational exposure to pesticides and Parkinson's disease: A systematic review and meta-analysis of cohort studies. *Environment International*, *46*, 30–43.
- Van Maele-Fabry, G., Lantin, A. C., Hoet, P., et al. (2011). Residential exposure to pesticides and childhood leukaemia: A systematic review and meta-analysis. *Environment International*, *37* (1), 280–291.
- Venkatapathy, R., & Wang, N. C. Y. (2013). Developmental toxicity prediction (Chapter 14). In B. Reisfeld & A. N. Mayeno (Eds.), *Computational toxicology: Volume II, methods in molecular biology* (Vol. 930, pp. 305–340). Heidelberg: Springer.
- Verma, J. P., Jaiswal, D. K., & Sagar, R. (2014). Pesticide relevance and their microbial degradation: A-state-of-art. *Reviews in Environmental Science & Biotechnology*, *13*, 429–466.
- Vighi, M., Masoero Garlanda, M., & Calamari, D. (1991). QSARs for toxicity of organophosphorus pesticides to *Daphnia* and honeybees. *Science of the Total Environment*, *109*(110), 605–622.
- Wan, Y. W., Tang, T. F., Zhou, Z. L. et al. (2009). Distribution and sources of organochlorine pesticides in Beijing Guanting Reservoir. *Journal of Ecology and Rural Environment*, *25*(1), 53–56, 68.
- Wang, J., Kliks, M. M., Jun, S., et al. (2010). Residues of organochlorine pesticides in honeys from different geographic regions. *Food Research International*, *43*, 2329–2334.
- Wang, J. Y., Yu, X. W., & Fang, L. (2014). Organochlorine pesticide content and distribution in coastal seafoods in Zhoushan, Zhejiang Province. *Marine Pollution Bulletin*, *80*, 288–292.
- Wang, B., Zhao, J. S., Yu, Y. J., et al. (2004). Quantitative structure-activity relationships and joint toxicity of substituted biphenyls]. *Huan Jing Ke Xue*, *25*(3), 89–93.
- Wigle, D. T., Turner, M. C., & Krewski, D. (2009). A systematic review and meta-analysis of childhood leukemia and parental occupational pesticide exposure. *Environmental Health Perspectives*, *117*, 1505–1513.
- Wu, H., Bertrand, K. A., Choi, A. L., et al. (2013a). Persistent organic pollutants and type 2 diabetes: A prospective analysis in the nurses' health study and meta-analysis. *Environmental Health Perspectives*, *121*, 153–161.
- Wu, C., Luo, Y., Gui, T., et al. (2014). Concentrations and potential health hazards of organochlorine pesticides in shallow groundwater of Taihu Lake region, China. *Science of the Total Environment*, *470–471*, 1047–1055.
- Wu, W. J., Qin, N., Zhu, Y., et al. (2013b). The residual levels and health risks of hexachlorocyclohexanes (HCHs) and dichloro-diphenyl-trichloroethanes (DDTs) in the fish from Lake Baiyangdian, North China. *Environmental Science and Pollution Research*, *20*, 5950–5962.
- Xu, M., Qiu, Y., Bignert, A., et al. (2015). Organochlorines in free-range hen and duck eggs from Shanghai: Occurrence and risk assessment. *Environmental Science and Pollution Research*, *22*, 1742–1749.
- Yao, Y., Harner, T., Blanchard, P., et al. (2008). Pesticides in the atmosphere across Canadian agricultural regions. *Environmental Science and Technology*, *42*, 5931–5937.
- Yuan, Y., Chen, C., Zheng, C., et al. (2014). Residue of chlorpyrifos and cypermethrin in vegetables and probabilistic exposure assessment for consumers in Zhejiang Province, China. *Food Control*, *36*, 63–68.
- Yusà, V., Coscollà, C., & Millet, M. (2014). New screening approach for risk assessment of pesticides in ambient air. *Atmospheric Environment*, *96*, 322–330.
- Zahouily, M., Rhihil, A., Bazoui, H., et al. (2002). Structure-toxicity relationships study of a series of organophosphorus insecticides. *Journal of Molecular Modeling*, *8*(5), 168–172.
- Zakarya, D., Boulaamail, A., Larfaoui, E. M., et al. (1997). QSARs for toxicity of DDT-type analogs using neural network. *SAR and QSAR in Environmental Research*, *6*, 183–203.
- Zakarya, D., Larfaoui, E. M., Boulaamail, A., et al. (1996). Analysis of structure-toxicity relationships for a series of amide herbicides using statistical methods and neural network. *SAR and QSAR in Environmental Research*, *5*(4), 269–279.

- Zakharov, A., & Lagunin, A. (2014). Computational toxicology in drug discovery: Opportunities and limitations (Chapter 11). In: L. Gorb et al. (Ed.), *Application of computational techniques in pharmacy and medicine, challenges and advances in computational chemistry and physics* (Vol. 17, pp. 325–367). Springer.
- Zhang, L., Dong, L., Shi, S., et al. (2009). Organochlorine pesticides contamination in surface soils from two pesticide factories in Southeast China. *Chemosphere*, *77*, 628–633.
- Zhang, L., Dong, L., Yang, W., et al. (2013). Passive air sampling of organochlorine pesticides and polychlorinated biphenyls in the Yangtze River Delta, China: Concentrations, distributions, and cancer risk assessment. *Environmental Pollution*, *181*, 159–166.
- Zhang, W. J., Jiang, F. B., & Ou, J. F. (2011). Global pesticide consumption and pollution: With China as a focus. *Proceedings of the International Academy of Ecology and Environmental Science*, *1*(2), 125–144.
- Zhang, H., Luo, Y., Zhao, Q., et al. (2006). Residues of organochlorine pesticides in Hong Kong soils. *Chemosphere*, *63*, 633–641.
- Zhang, X., Wu, M., Yao, H., et al. (2016). Pesticide poisoning and neurobehavioral function among farm workers in Jiangsu, People's Republic of China. *Cortex*, *74*, 396–404.
- Zhao, L., Hou, H., & Guo, P. Y. (2009). Distribution of organochlorine pesticides in soils in Haihe River and Haihe estuary area. *China Environmental Science*, *30*(2), 543–550.
- Zheng, S., Chen, B., Qiu, X., et al. (2016). Distribution and risk assessment of 82 pesticides in Jiulong River and estuary in South China. *Chemosphere*, *144*, 1177–1192.
- Zhou, Q., Sun, X., Gao, R., et al. (2010). Mechanism study on OH-initiated atmospheric degradation of the organophosphorus pesticide chlorpyrifos. *Journal of Molecular Structure THEOCHEM*, *952*, 8–15.
- Zhu, H., Martin, T. M., Ye, L., et al. (2009a). Quantitative structure—Activity relationship modeling of rat acute toxicity by oral exposure. *Chemical Research in Toxicology*, *22*(12), 1913–1921.
- Zhu, H., Ye, L., Richard, A., et al. (2009b). A novel two-step hierarchical quantitative structure-activity relationship modeling work flow for predicting acute toxicity of chemicals in rodents. *Environmental Health Perspectives*, *117*(8), 1257–1264.
- Zvinavashe, E., Du, T., Griff, T., et al. (2009). Quantitative structure-activity relationship modeling of the toxicity of Organothiophosphate pesticides to *Daphnia magna* and *Cyprinus carpio*. *Chemosphere*, *75*(11), 1531–1538. doi:[10.1016/j.chemosphere.2009.01.081](https://doi.org/10.1016/j.chemosphere.2009.01.081).

Counter-Propagation Artificial Neural Network Models for Prediction of Carcinogenicity of Non-congeneric Chemicals for Regulatory Uses

N. Fjodorova, M. Novic, S. Zuperl and K. Venko

Abstract The evaluation of carcinogenic hazard of chemicals to human is nowadays one of the most challenging tasks. Quantitative structure–activity relationship (QSAR) models are welcome tools to cope with complex, expensive and time consuming experimental methods for evaluation of carcinogenic potency. Therefore, in last decade, vast effort was involved to introduce new in silico models for prediction of carcinogenicity of non-congeneric chemicals that can be effectively used for regulatory purposes in the scope of new legislation REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals). In this chapter we focus on models developed in the scope of CAESAR and PROSIL projects which were implemented in on-line available internet platform VEGA (<http://www.vega-qsar.eu/use-qsar.html>). These QSAR models for prediction of carcinogenic potency are based on counter propagation artificial neural network algorithm (CPANN). CP ANN algorithm represents a suitable tool for modeling of complex biological data like carcinogenicity. We emphasized on the representation of key development steps needed to be involved in model construction to meet requirement of five OECD principles. First of all, it reported the description of carcinogenicity endpoint and analysis of quality of chemical and biological data (principle 1), followed by an explanation of the CPANN algorithm selected for modelling (principle 2). Next, the interpretation of domain of applicability for non-congeneric chemicals (principle 3) was given. Furthermore, the statistical performance characteristics of models in

N. Fjodorova (✉) · M. Novic · S. Zuperl · K. Venko
National Institute of Chemistry, Hajdrihova 19, 1001 Ljubljana, Slovenia
e-mail: Natalja.Fjodorova@ki.si

M. Novic
e-mail: marjana.novic@ki.si

S. Zuperl
e-mail: spela.zuperl@ki.si

K. Venko
e-mail: katja.venko@ki.si

sense of its goodness-of-fit, robustness and predictivity was reported (principle 4), and finally, the mechanistic interpretation of models on the basis of selected types of descriptors and structural alerts of studied chemicals was represented (principle 5).

Keywords Counter propagation artificial network (CPANN) • In silico models • (Quantitative) structure-activity relationship ((Q)SAR) • Carcinogenicity • Models for regulatory use • Non-congeneric chemicals • REACH

Abbreviations

AC	Accuracy
AD	Applicability domain
CAESAR	Computer Assisted Evaluation of industrial chemical Substances According to Regulation
CPANN	Counter propagation artificial neural network
CPDB	Carcinogenic Potency Database
ECHA	European Chemical Agency
ED	Euclidean distances
FN	False negatives
FP	False positives
IARC	International Agency for Research of Cancer
IATA	Integrated Approach to Testing and Assessment
NP	Not positive
P	Positive
QSAR	Quantitative structure–activity relationship
REACH	Registration, Evaluation, Authorization and Restriction of Chemicals
SA	Structural alerts
SAR	Structure–activity relationship
TD	Tumourgenic dose
TN	True negatives
TP	True positives

1 Introduction

The carcinogenic potency of chemicals is of great importance in assessment of human health safety. The experimental carcinogenicity tests are expensive and require animal testing, which is contrary to the policy in EU member states to replace, reduce and refine the use of animals in science, (the so called 3Rs policy). The implementation of REACH, which aims to fill information gaps for large number of chemicals in order to minimize animal testing, initiated the employment of in silico models for safety assessment of chemicals EC (2008) including the (Quantitative)Structure-Activity Relationships (Q)SAR approach.

A thorough analysis of the use of *in silico* information in a regulatory setting shows that the number of REACH dossiers for which read-across and/or (Q)SAR was used to replace experimental evidence hovers around 30% (ECHA 2014).

The more detailed information of QSAR models and software tools for predicting genotoxicity and carcinogenicity was published as JRS Scientific and Technical Reports by Serafimova et al. (2010). The following databases are used for QSAR modeling of genotoxicity and carcinogenicity: CPDB, Danish QSAR database, DSSTOX, ECHA CHEM, ESIS, EXCHEM, GAP, IARC, ISSCAN, NTP, ToxRefDB, TOXNET, GENE-TOX. The evaluation of carcinogenic potency started with so called structure-activity relationships (SAR) method. At first, the nineteen (19) structural alerts (SA) for carcinogenicity was proposed by Ashby (1985), then thirty three (33) SAs were offered by Bailey et al. (2005) and twenty nine (29) SAs was submitted by Kazius et al. (2005) and, finally, the thirty three (33) SAs were developed by Benigni and Bossa (2008a) and incorporated into Tox Tree expert system by Benigni et al. (2008b). It should be noted that SAR models only use (sub) structure information and can therefore be seen as a more formal way of performing read-across with a given reference set of data, as the property of one or more substances is directly used to predict the property of the substance of interest Jacobs et al. (2016).

The most widespread approaches used in carcinogenicity models are rule-based, statistical and hybrid (Serafimova et al. 2010).

The most well-known public or/and commercial software for genotoxicity and carcinogenicity with indication of web sources and authors are represented in the Table 1.

The needs of industry and regulators to assess thousands of compounds initiate the development of high-throughput assays combined with innovative data-mining and *in silico* methods. Various initiatives in this regard have begun, including CAESAR, OSIRIS, CHEMOMENTUM, CHEMPREDICT, OpenTox, EPAA, and ToxCast™ (Benfenati et al. 2009).

European Chemical Agency (ECHA) declared that non-testing methods to assess carcinogenic hazard to humans include the Quantitative Structure-Activity Relationships ((Q)SARs) as well as chemical grouping for read across approaches ECHA (2014). The QSAR models for prediction of toxicological properties of substances take into account the quantitative parameters describing the structure as well as physico-chemical (reactivity) properties of considered substances. SAR models only use the (sub) structures information (fragments), therefore they can be considered as a more formal way of performing read across with a given reference set of data. The property of one or more substances here is directly used to predict the property of a new substance. It should be highlighted that QSAR method is based on a large good quality dataset using the robust scientific and statistical concept. Read across in some sense is more subjective approach, but it can provide the more specific information. In the latest article by Jacobs et al. (2016), the Integrated Approach to Testing and Assessment (IATA) was proposed. The authors supposed that in future a combination of (Q)SAR with read across (local validity analysis of models), may be of greater reliability for decision making. VEGA

Table 1 The list of software for genotoxicity and carcinogenicity

Name of software	Web source	References and/or organization
CAESAR (VEGA platform)	(http://www.caesar-project.eu/);	Fjodorova et al. (2010a)
HazardExpert	(http://www.compudrug.com)	Lewis et al. (2002)
Lazar	http://lazar.in-silico.de	Helma (2006)
MDL-QSAR	http://www.symyx.com	Contrera et al. (2005a) and Valerio et al. (2007)
Multicase (MCASE/MC4PC) MultiCASE Inc.	http://www.multicase.com	Matthews and Contrera (1998) and Matthews et al. (2006a, b)
OncoLogic™	http://www.epa.gov/oppt/newchems/tools/oncologic.htm	Woo and Lai (2005)
TOPKAT (Accelrys)	http://www.accelrys.com	Enslein et al. (1994) and Prival (2001)
Toxtree	http://ecb.jrc.ec.europa.eu/qsar/	Benigni et al. (2010) and Benigni et al. (2009)
OECD toolbox	http://toolbox.oasis-lmc.org	Benigni et al. (2008b), Serafimova et al. (2007)
PASS	http://www.way2drug.com/ http://www.way2drug.com/PASSOnline/downloads.php	Poroikov et al. (2010)
GAP—genetic activity profile database	http://www.ils-inc.com	Initially developed by US EPA and IARC, and now by ILS
Derek	https://www.lhasalimited.org/products/derek-nexus.htm	Lhasa Ltd., Marchant (1996)
MolCode toolbox	http://www.chemistry-software.com/molcode/index.html	Molcode Ltd.

platform represents the example of application QSAR and read across methods. VEGA was developed and accepted for regulatory use (Jacobs et al. 2016).

In this chapter we discussed the main features of models employed in VEGA platform (Vega web site: <http://www.vega-qsar.eu/>) developed within European Commission (EC) funded project CAESAR (Computer Assisted Evaluation of industrial chemical Substances According to Regulation) (CAESAR web site: <http://www.caesar-project.eu>) and ongoing project PROSIL (PROSIL web site: <http://www.life-prosil.eu/>).

The categorical or qualitative models for prediction of carcinogenic potency of non-congeneric chemicals using Counter Propagation Artificial Neural Network (CPANN) method were presented in this chapter. These models have been developed in accordance with 5 principles of validation (Q)SAR models for their use in regulatory assessment of chemical safety adopted by OECD member countries in November 2004 OECD (2004a). A full report from the OECD Expert

Group on (Q)SARs was also published in 2004: OECD (2004b). In February 2007, the OECD published a “Guidance Document on the Validation of (Q)SAR Models” (OECD 2007).

Within CAESAR project, the data mining approach has been improved using a highly verified set of compounds (all chemical structures have been double-checked, and experimental data verified in case of some unusual finding, compared to similar compounds). A wide series of chemical descriptors were adopted. Different algorithms have been developed, this resulted in a series of models. The best performance was obtained in the case of CPANN algorithm which is reported here. The predictive power of models is one of the most important characteristics in QSAR modeling. Benigni et al. (2008c) pointed out that the prediction reliability should be checked by means of an external test set with new chemicals not used in modeling. It was stressed that the models for regulatory purposes should be connected with high sensitivity, i.e., the ability to correctly identify true positives. In the CAESAR project, an external dataset of 738 chemicals was composed and external validation of models was done. It was shown how one can increase the number of correctly predicted carcinogens using correlation between threshold of categorical models and sensitivity and specificity. It was demonstrated how threshold influences overall performance of models.

Preliminary results of carcinogenicity modeling using CPANN algorithm obtained in the scope of CAESAR project are described in an article by Fjodorova et al. (2010a), while final results were reported in the other articles of Fjodorova et al. (2010b, c, 2012). The differences in carcinogenic potency obtained in CAESAR CPANN models and Toxtree expert system were discussed in the article Fjodorova et al. (2014).

The main advantage of neural network modeling is that the complex, non-linear relationships can be modeled without any assumptions about the form of the model. Large datasets can be examined. Vračko et al. (2004) described why the neural networks are able to cope with noisy data and are fault-tolerant.

The CPANN models were incorporated in the VEGA platform and could serve for the preliminary ranking and prioritization of chemicals for carcinogenic potency, as required by REACH.

2 The Determination of Endpoint and Quality of Chemical and Biological Data (Principle 1)

2.1 The Criteria for Cancer Risk Assessment and Test Guidelines

International Agency for Research of Cancer (IARC 2016) established the criteria for Cancer Risk Assessment in the IARC Monographs. According to these criteria, chemicals can be classified as *Carcinogenic to humans* (Group 1), *Probably*

carcinogenic to humans (Group 2A), *Possibly carcinogenic to humans* (Group 2B), *Not classifiable as to its carcinogenicity to humans* (Group 3) and *Probably not carcinogenic to humans* (Group 4)

Carcinogenic potency can be assigned by studies in human, in experimental animals as well as using mechanistic and other relevant data.

Chronic oral toxicity and carcinogenicity tests are described in “OECD Environment, Health and Safety Publications Series on Testing and Assessment №35 Guidance Notes for Analysis and Evaluation of Chronic Toxicity and Carcinogenicity Studies (OECD 2002). Additionally, description of the Chronic Toxicity and Carcinogenicity Studies is given in OECD Test Guidelines 451–453. The original Test Guideline 453 for Combined Chronic Toxicity/Carcinogenicity Studies was adopted in 1981. It includes the carcinogenicity hazard testing and assessment as described in OECD Guidelines TG 453 available as OECD TG 453 (2009) and the test guideline TG 451 on Carcinogenicity Studies available as OECD TG 451 (2009). OECD Guidelines for the Testing of Chemicals (TGs) are periodically reviewed in the light of scientific progress, changing assessment practices and animal welfare considerations.

The majority of carcinogenicity studies are carried out in rodent species, and this Test Guideline is intended therefore to apply primarily to studies carried out in these species.

2.2 The Carcinogenicity Endpoint Used in VEGA Models

In the CAESAR and PROSUL models, the carcinogenic potency for rats was selected as response because such data in risk assessment (Combes et al. 2008) are often considered to be more suitable for human carcinogenicity prediction. The term “carcinogen” generally refers to an agent, mixture, or exposure that increases the age-specific incidence of cancer. Carcinogen identification is an activity grounded in the evaluation of the results of scientific research. Tumourgenic dose is accepted for characterization of carcinogenicity. The tumourgenic dose TD₅₀ used in our study is defined as the tumourgenic dose rate where 50% of the test animals got any kind of cancer. In other words, the TD₅₀ is that chronic dose rate (in mg/kg body weight/day or mmol/kg body weight/day (mmol/kg-bw/day)) which would give half of animal tumors within some standard experiment time, the “standard life span” for the species (Peto et al. 1984).

An assignment of carcinogenic categorical activity based on evidence for or against activity within the species group in Target Sites of Rats (Male, Female or Both) has been accepted. Hence, “active” or positive (P) or carcinogen was assigned for a compound if one or more TD₅₀ and the tumor site are listed for one or more rat carcinogenicity sex/species cell (rat male, rat female, rat both) and “inactive” or not positive (NP) or non- carcinogens was assigned for a compound if no TD₅₀ or tumor site are listed and one or more “no positive results” entry for one or more rat carcinogenicity sex/species cell, i.e., one or more experiments are

reported in the Carcinogenic Potency Database (CPDB) (<https://toxnet.nlm.nih.gov/cpdb/cpdb.html>) for species, but none are positive. In other words, chemicals were classified as not carcinogenic when the results obtained during all animal tests on rats were assigned as not positive (NP) (or not active) and in contrary, compounds were classified as positive (P) (or active) when any of the in vivo assays gave a well-defined TD₅₀ value.

In the CAESAR model, the studied database contained 805 compounds. Among them the 421 chemicals were classified as carcinogenic (P) and remaining 384 as non-carcinogens (NP). In PROSIL model, the dataset of 792 chemicals was used. Among these chemicals, 609 compounds were carcinogens and 185 chemicals were non-carcinogens.

2.3 Quality of Chemical and Biological Data Used in the Model

The chemicals involved in the study belong to different chemical classes (including halogenated hydrocarbons, aromatic compounds, ketones, aldehydes, organic acids, heterocyclic and polycyclic compounds, amines, amides, sulfonates, etc.), so called non-congeneric substances. The work in CAESAR and PROSIL projects was addressed to industrial chemicals, referring to the REACH initiative. The aim was to cover chemical space as much as possible. In the scope of CAESAR project, the initial dataset of 1481 chemicals was taken from Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network described in the articles by Richard and Williams (2001), Richard (2004). DSSToxPublic Database was built from the Lois Gold Carcinogenic Database (CPDB), CPDB (<https://toxnet.nlm.nih.gov/cpdb/cpdb.html>).

The data used in QSAR modelling should be homogeneous to get reliable models. Moreover, for some structures it is not possible to transform them into descriptors, therefore they should be eliminated from the model. Hence, the initial dataset in CAESAR project has been cleaned of all incorrect structures, ambiguous or mixed structures, polymers, inorganic compounds, metallo-organic compounds, salts, complexes and compounds without well-defined structure. The obtained data and structures of chemicals were cross-checked by at least two partners.

The final data set was composed of 805 chemicals in CAESAR project and 792 chemicals in PROSIL project.

The detailed information about the dataset used in the carcinogenicity modeling in CAESAR project was published in the article (Fjodorova and Novič 2013).

It should be highlighted that data used in the study were obtained from standard protocols and meet requirements for QSAR modeling.

2.4 Training, Test and External Validation Sets

To prepare data for modelling, the dataset of 805 chemicals was subdivided into training (644 chemicals) and test (161 chemicals) sets using the sub-sorting of chemicals according to functional groups and following procedure aimed to distinguish between connectivity aspects. This part of study has been done in the Helmholtz Centre for Environmental Research-UFZ in Germany by a partner in the CAESAR project. The sorting of the compounds pointed here is implemented in the software system ChemProp (Schüürmann et al. 1997; Schüürmann et al. 2007).

Additional 738 chemicals different from those in the data set of 805 compounds were used as the external validation set, being described by the same type of structural descriptors as employed in our model. To assess predictive abilities of the selected CAESAR model, a commercial database has been queried to extract new chemical compounds to be tested. Leadscope software allows accessing some QSAR ready database and the “FDA 2009 SAR Carcinogenicity—SAR Structures” database consisting of 2090 compounds has been extracted from the Leadscope environment in terms of structure information and carcinogenic activity label (based on different mammalian species) and compared with the CAESAR dataset of 805 compounds (ChemFinder Ultra 10.0 2009).

The two databases were prepared in the form of SDF files and specific check to search for duplicates has been performed. The compounds in common between the two sources were analysed to verify consistency in the experimental carcinogenicity class assigned by the two sources.

A total of 655 compounds were in common and for them the CAESAR assignment was compared with the Leadscope one. The assignment of toxicity class for Leadscope chemicals was based on rat data only and chemicals have been classified as carcinogens if at least one of the two genders (male or female rat) was labelled in Leadscope as positive or intermediate level carcinogen.

Based on this group of 655 compounds, the concordance of the two assignments was of 367 positive chemicals and 257 non-carcinogenic ones. Only 31 compounds were classified differently (11 positive in CAESAR dataset but negative for Leadscope and 20 in the opposite situation); hence the overall concordance was above 95%.

Since the concordance between the two experimental sources is very high, the Leadscope database was considered as a reliable source of new compounds to test the CAESAR model.

After exclusion of those chemicals already present in the CAESAR dataset, it was possible to select as an external test set 738 compounds with experimental data on rats. The external test compounds have been submitted to the CAESAR model to obtain the predicted power of model.

3 CPANN Algorithm (Principle 2)

The CPANN method was used in modelling; it belongs to self-organizing map technique that is often used to analyse the data in multi-dimensional space. The basis of this technique is a non-linear projection from multi-dimensional space onto a two-dimensional map. The topology preserving projection is achieved via a non-linear algorithm known as training. The fundamental property of the trained network is close vicinity of similar objects. Therefore, it is expected that chemicals with similar structure will form the clusters, which is the case of examination.

The architecture of CPANN is shown in Fig. 1.

The network constructed of neurons has two layers: input layer (Kohonen layer) containing encoded information of structure expressed as descriptors values and output layer (response). Both layers of neurons are placed exactly one above the other and the output layer has exactly the same layout of neurons as the input one (Zupan et al. 1997).

The input layer has a number of levels (weights of the input neurons corresponding to the number of descriptors, i.e., the dimension of input vector X), while the output layer has as many levels as the target vectors have responses.

Kohonen maps enable visualisation of the distribution of chemicals (in the top map) and distribution of descriptors values (in weight levels maps). CPANN, in turn, is a generalization of self-organizing map. Additionally, it takes into account the property (output) values (Vračko et al. 1999, 2004) and is encompassed in the output layer. The learning in the input layer in the CPANN is the same as in Kohonen neural network, i.e., the similarity among input variables determines the arrangement of objects in the input layer map. When the arrangement is set, the

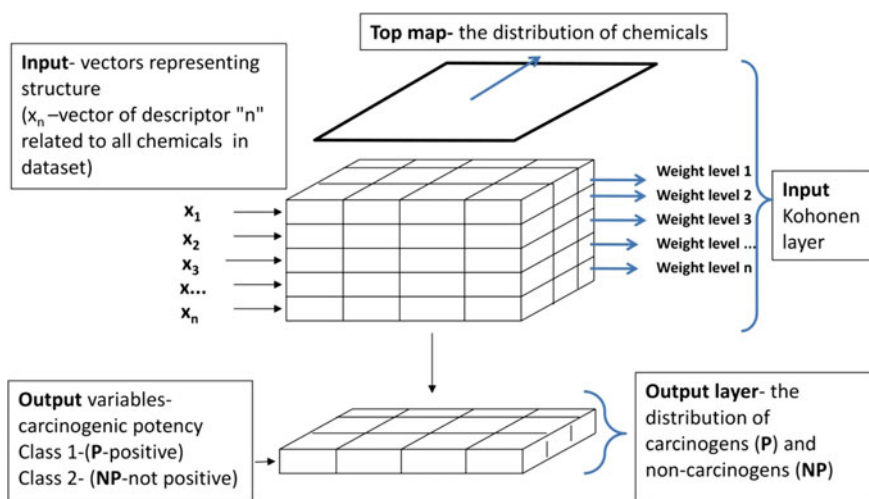


Fig. 1 The architecture of CPANN

positions of objects are projected to the output where the weights are modified in a way that the weights on projected positions are getting similar to the values of corresponding objects.

The model for prediction of carcinogenic class (P/NP) is represented in the study.

In Fig. 1, the inputs $x_1, x_2, x_3, \dots, x_n$ are vector components representing chemical structure which corresponds to descriptors calculated for all chemicals used in training dataset. In the other words, $x_{1i}, x_{2i}, x_{3i}, \dots, x_{ni}$ can be represented as a matrix of descriptors 1, 2, 3, ..., n values for all chemicals in the training dataset, respectively. The distribution of chemicals and their clusters in 2D space is examined in the Kohonen top map. Weight levels 1, 2, 3, ..., n are the maps with distribution of particular descriptors 1, 2, 3, ..., n, correspondingly.

The output layer is associated with the output values so called target $T_n = (t_{n1}, t_{n2}, \dots, t_{nj}, \dots, t_{np})$ which is a p -component vector of zeros and ones (for classification model). One dimensional target in our classification models expresses carcinogenicity class (P-positive = 1 and NP-not positive = 0). The neural network is trained to respond for each input structure representation X_n from the training set with the output vector Out_n identical to the target (class-vector) T_n . Thus, the output variables in Fig. 1 are expressed in the output layer as a carcinogenicity class (class 1 was marked as positive (P)-carcinogen and class 2 as non-positive (NP)-non-carcinogen).

The Kohonen input layer of the CP ANN consists of $n_x \times n_y$ neurons. After the learning, the objects are organized in such a way that similar objects are situated close to each other. It is to emphasize that only the input values participate in this phase of learning (unsupervised step). For this step, no knowledge about the target vector is needed (Zupan et al. 1997).

In the second step the positions of objects are projected to the output layer, where the weights are adjusted to output values (supervised step). The trained output layer consists of $n_x \times n_y$ output neurons arranged in squared neighborhood. After the training, each weight of the output neurons out_j is a real number between 0.0 and 1.0. For the final prediction of classes, the response surface values must be again transformed into discrete values, 0 and 1. The threshold value between 0.01 and 0.99 must be determined for each class.

More detailed description of CPANN can be found in the literature (Zupan and Gasteiger 1999; Zupan et al. 1997; Mazzatorta et al. 2003).

4 Domain of Applicability (Principle 3)

The definition of the applicability domain (AD) of a QSAR model is very useful to define boundaries whereby the obtained predicted values can be trusted with confidence. So far no standard solutions have been agreed within the scientific community to optimally define these boundaries, but often the proposed solutions rely

on chemometrics methods. The state of art of methods for identifying the domain of applicability of (Q)SARs is given in paper (Netzeva et al. 2005).

Amongst the model builders there are different interpretations of the definition of the term “applicability domain” (VEGA, TOPKAT, MultiCASE all have their own definitions) (Jacobs et al. 2016).

For carcinogenicity endpoint, CAESAR implemented a tool for the general evaluation of the AD based on the descriptor range for the dataset. Therefore, predicted values for chemicals outside the descriptor range can be judged as less reliable. Though, this kind of estimation of the AD does not address two key aspects. Firstly, the chemical space characterised by the descriptor range does not take into account the density of compounds distribution, so it might happen that the target chemical falls in an area poorly represented in the training set. Moreover, since the AD relies on the chemical descriptors alone, the output layer (the property under investigation) is neglected. To overcome these aspects, CAESAR developed a further tool for the AD assessment, based on the measurement, through a similarity score, of the six most similar chemicals in the training set. It can be used to evaluate if these compounds are really representative for the unknown compound. Furthermore, a visualisation of these compounds is offered, which can be used to independently evaluate the compounds. Finally, a quantitative report of the error between the observed and predicted activity is also provided for these substances, so that it is possible to argue about wrong behaviour for the model in the chemical area that better represents the compound of interest. This feature was incorporated in the VEGA platform.

Because the CPANN models were employed in the study, the future search for the characterization of domain of applicability for this kind of models was performed. As an outcome, the new metrics for the evaluation of an AD for the non-linear models (like neural networks) for the diverse set of chemicals was proposed and described in the articles (Fjodorova et al. 2011; Minovski et al. 2013). The authors proposed to use the Euclidean distances (ED) between an object (molecule) and the corresponding excited neuron of the neural network (Fjodorova et al. 2011; Minovski et al. 2013) and between an object (molecule) and the representative object (vector of average values of descriptors) (Fjodorova et al. 2011).

The ED between objects (molecules) and central neuron in Kohonen layer of CP ANN models is the essential characteristic of neural network. This metric was used to compare the training and test sets chemical coverage of models with respect to false predicted chemical space.

The ED represents the interval between a node in the Kohonen layer and an input pattern. The distances are unitless, because all descriptors have been auto scaled. It was noted in the literature (Maran et al. 2004) that in fact, the sum of distances of all molecules obtained after training of the network during one epoch is equal to the cumulative training error associated with the Kohonen layer. Thus, the ED depends on the input values of descriptors from one side and is connected with the errors from the other.

If the ED between objects (molecules) and central neuron in Kohonen layer of CPANN models gave us ability to compare the training and test sets chemical coverage of models with respect to false predicted chemical space, the ED between vectors of real values of descriptors and the vector of average values of descriptors was used to explore the coverage of the descriptor space for the training and the test set chemicals in the models with respect to the space of wrongly predicted chemicals.

The ED in the non-linear models demonstrates boundaries where the model was built and is applicable with the determined reliability.

5 Statistical Performance Characteristics of Model (Goodness-of-Fit, Robustness and Predictivity) (Principle 4)

5.1 Parameters Used for Evaluation of Classification Models

A common way to evaluate the performance of classification models (or classifiers) is to employ a confusion matrix (see Table 2) according to the method of Cooper et al. (1979).

In the confusion matrix the four different possible outcomes of a single prediction for two-class problem are displayed. The rows represent the number of entries belonging to actual (observed) class, while the columns represent the entries belonging to predicted class. N_{negative} and N_{positive} are the number of negative (non-carcinogens) and positive compounds (carcinogens) in the dataset. TP denotes

Table 2 Confusion matrix for two class classifier (P-positive and N-negative)

		Predicted		
		Non-carcinogens (Negative)	Carcinogens (Positive)	Total predicted
Observed	Non-carcinogens (Negative)	TN	FP	$N_{\text{negative}} = \text{TN} + \text{FP}$
	Carcinogens (Positive)	FN	TP	$N_{\text{positive}} = \text{FN} + \text{TP}$
	Total observed	$\text{TN} + \text{FN}$	$\text{FP} + \text{TP}$	$N_{\text{total}} = N_{\text{negative}} + N_{\text{positive}}$

*Definitions in Table 1

TP-True positive

TN-True negative

FP-False positive

FN-False negative

N_{negative} is the number of negative (non-carcinogens) in the dataset

N_{positive} is the number of positive compounds (carcinogens) in the dataset

N_{total} is total number of negative (non-carcinogens) and positive compounds (carcinogens) in the dataset

the number of true positives, and TN denotes the number of true negatives. FP (false positives) is the number of errors made by predicting a compound of being active (carcinogen) while it is not; FN (false negatives) is the number of incorrectly predicted negatives (non-carcinogens).

Cooper statistics express the ability of classification models to detect known active compounds (sensitivity), non-active compounds (specificity), and all chemicals in general (accuracy). See Eqs. (1)–(3).

The main classification parameter is the *accuracy* (AC) (or concordance). It is determined using the equation:

$$AC = \frac{TN + TP}{TN + FN + FP + TP} \quad (1)$$

AC is defined as the total number of non-carcinogens and carcinogens correctly predicted among the total number of compounds.

The others statistical parameters of interest are *sensitivity*, *specificity*, *positive predictivity*, *negative predictivity*, *false negative rate*, *false positive rate* and etc. *Sensitivity* is defined as the percentage of correctly classified carcinogens among the total number of carcinogens. It can be determined as the *true positive rate* and can be expressed as follows:

$$TP \text{ rate} = \frac{TP}{TP + FN} = \textit{Sensitivity} \quad (2)$$

Specificity shows the percentage of correctly classified non-carcinogens among the total number of non-carcinogens and relates to *true negative rate*. The following equation corresponds to specificity:

$$TN \text{ rate} = \frac{TN}{TN + FP} = \textit{Specificity} \quad (3)$$

Training and test sets were composed for evaluation of models. *Training set* represents class values for learning. *Test set* represents class values for evaluation. Hypothesis is used to establish classification in the test set, which is compared to known one.

5.2 Internal Validation

For evaluation of *goodness-of-fit* or robustness of CAESAR models, the internal performance of model based on the training set (644 compounds) was applied. Several diagnostic statistical tools were implemented for characterization the *goodness-of-prediction* or predictability of the obtained models. Firstly, statistical performance of the test set (161 compounds) was calculated. Secondly, internal cross-validation (Eriksson et al. 1996, 2003) (CV) using “leave 20% out” test was

done. It was performed on a training set of 644 compounds, so that the set was divided into five training sets, each containing 80% of compounds, and five test sets with 20% of compounds. The sets were selected randomly in a way that each compound was exactly one time a part of the test set and four times a part of the training set.

5.3 External Validation

External validation is commonly used for the predictivity and reliability of QSAR models (Perkins et al. 2003; Golbraikh and Tropsha 2002).

Therefore, the predictive performance of QSAR models should be evaluated using a validation set of compounds that were not used to generate the model. The validation set of 738 compounds was provided by the CAESAR project partner (*Istituto di Ricerche Farmacologiche “Mario Negri” (IRFMN), Milano, Italy*) and implemented for validation of models.

In conclusion, it should be highlighted that the evaluation of the classification system was done using the so-called internal training set (644 compounds) and test set (161 compounds), cross validation using leave-20%-out test, and external validation test set (738 compounds). The external test set included chemicals that were not considered in the modeling. The results of different CAESAR models are described below.

5.4 Model Using Eight MDL Descriptors

With 8 MDL descriptors and CP ANN algorithms described above, dozens of models were produced. After their evaluation, one model was accepted as the best one (Model A). The statistical performance of this model is presented in Table 3. The Cooper statistics based on the training set indicated an accuracy of 91%, high value of sensitivity (96%) and specificity (86%). Again, for the test set (161 compounds), we obtained accuracy equal to 73%, sensitivity (75%) and specificity (69%). Cross validation (leave 20% out) results gave us accuracy 66%. From the

Table 3 Statistical performance of models using 8MDL descriptors (Model A) and 12 Dragon descriptors (Model B)

Internal validation (%)	Model A (8MDL descriptors)		Model B (12Dragon descriptors)	
	Training (644 compounds)	Test (161 compounds)	Training (644 compounds)	Test (161 compounds)
Accuracy	91	73	89	69
Sensitivity	96	75	90	75
Specificity	86	69	87	61

results of external validation (738 compounds), we have got an accuracy of 61.4%. The obtained results indicated that models possessed good stability. Reliability and robustness of model are high as we get good statistical performance for all criteria, on both the internal and external sets.

The model reliability should be connected with high sensitivity (correctly predicted carcinogens) to ensure public safety (Benigni et al. 2008c). In this chapter, the way to increase sensitivity by changing the threshold of a model is demonstrated. Figure 2 shows the accuracy, sensitivity and specificity for the test set for model A depending on the threshold. From a regulatory perspective, the higher sensitivity (correctly predicted carcinogens) in prediction of carcinogens is more desirable than high specificity (correctly predicted non-carcinogens). Changing the threshold in model A, one can vary sensitivity and specificity depending on the needs. In the interval of threshold from 0.05 to 0.9, the accuracy is greater than 60%. Setting on threshold to 0.05, it is possible to increase the sensitivity till 90% without considerable reduction of accuracy as it still remains at the level of 60%. On the other hand one should keep in mind that increasing of sensitivity leads to considerable reduction of specificity till approximately 20% in case of threshold 0.05 (see Fig. 2).

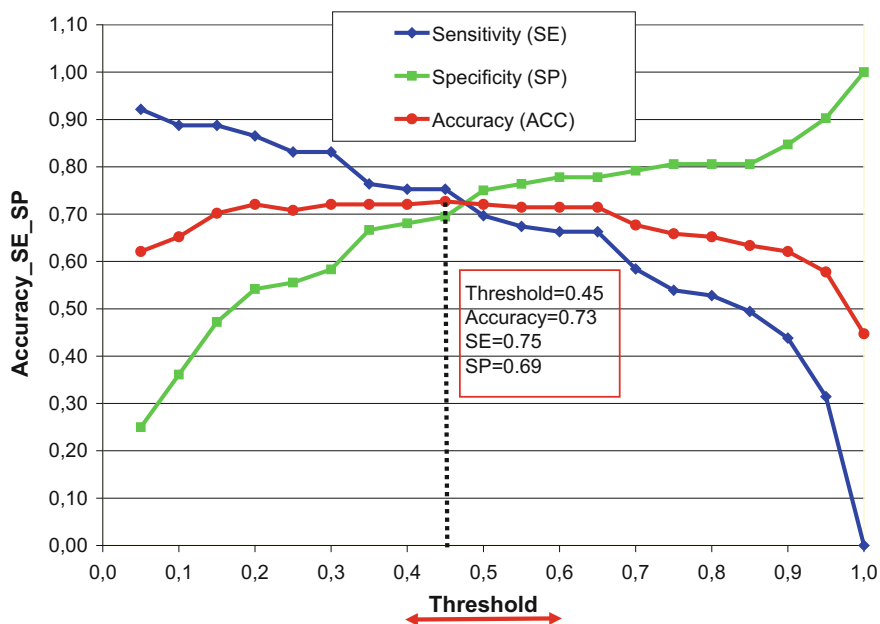


Fig. 2 Accuracy (ACC), sensitivity (SE) and specificity (SP) of test set (161 compounds) versus threshold for CP ANN model A

Another important feature of models for regulatory purposes is the reproducibility. Therefore the parameters of model have to be fixed. The user does not need to optimize the model parameters. In Fig. 2 the optimal model performance corresponding to the threshold equal to 0.45 is shown.

5.5 Model Using Twelve Dragon Descriptors

As an alternative choice, Dragon descriptors have been used for prediction of carcinogenicity using CPANN algorithm. 12 Dragon descriptors were employed in the model. The same dataset of 805 chemicals was used. An optimal model with dimension of neural network $35 * 35$ and number of learning epochs equal to 200 was selected. The threshold was set up at 0.5.

The Cooper statistics of the model with 12 Dragon descriptors based on the training set (644 compounds) indicated the accuracy 89%, sensitivity 90% and specificity 87%, while for the test set (161 compounds) the accuracy was equal to 69%, sensitivity 75% and specificity 61%. The threshold was set at 0.5. The characterization of this model B is given in the Table 1.

5.6 Validation of Models Using External Set of 738 Compounds

The CAESAR applet available on the intranet was used to predict a set of 738 compounds which were not used in modelling (not presented in the CAESAR dataset of 805 chemicals). These chemicals were provided with carcinogenicity class assigned on the basis of experiments and extracted from the Leadscape software. The predictions obtained for these compounds were summarized.

Overall, the performances obtained with this externally predicted dataset are as follows: accuracy = 61.4 and 60.0%; sensitivity = 64.0 and 61.8% and specificity = 58.9% and 58.4% respectively for the model with MDL Dragon descriptors.

6 Mechanistic Interpretation of Models (Principle 5)

According to five OECD principles for establishing the validity of quantitative structure-activity relationship (QSAR) models for use in regulatory assessment of chemical safety, a QSAR should be associated with a mechanistic interpretation, if possible. The intent of this principle is to consider the relationship between descriptors and an endpoint to find out a potential mechanism of action.

Development of a structure-information approach which is based on application of different structural descriptors including the electrotopological ones shows the new opportunities in prediction of biological activity and properties in contrast to the mechanism based approach (Hall 2004; Kier and Hall 2005; Hall and Hall 2005).

The models based on the above pointed approaches are established independent of explicit three-dimensional (3-D) structure information and are directly interpretable in terms of the implicit structure information (Rose and Hall 2003). The authors (Hall 2004) demonstrated a wide range of applicability of such models for relatively big datasets (e.g., for prediction of aqueous solubility, AMES mutagenicity, fish toxicity and others). In the case of carcinogenicity there are a variety of mechanisms and pathways, including genotoxic and epigenetic ones that might play a role in the observed toxic effect. The application of structure-information approach which is “mechanism-free” makes our task simpler and thus feasible because it is not necessary to assume various mechanistic steps in order to make computations for such complicated biological property like carcinogenicity. This method is free of approximations and computations related to assumed mechanism of interaction. This aspect is very important especially for modelling carcinogenicity using non-congeneric set of substances and aimed for prediction of a wide diversity of chemicals.

The MDL and Dragon chemical descriptors selected within the CAESAR models are presented in Tables 4 and 5, correspondingly.

MDL descriptors (see Table 4) contain electrotopological E-state, connectivity and others descriptors. E-state indices are a combination of electronic, topological and valence state information. These indices incorporate information related to atom, types and electron accessibility, hydrogen atom E-states, and connectivities that are influenced by all of the sub-structural features of a molecule (Kier and Hall 1999a, b, 2001).

Elements identity and skeletal connection contain structure information while valence state definition includes relationship for valence state electronegativity and atom/group molar volume. Based on these important features of molecules, together

Table 4 Eight MDL descriptors employed in CAESAR model

MDL_ID descriptor code	Symbol	Definition
<i>MDL005</i>	<i>SdsCH</i>	Sum of all (= CH –) E-State values in molecule
<i>MDL051</i>	<i>SdssC_acnt</i>	Count of all (= C <) groups in molecule
<i>MDL062</i>	<i>SdsN_acnt</i>	Count of all (= N) groups in molecule
<i>MDL114</i>	<i>dxp9</i>	Difference simple 9th order path chi indices
<i>MDL130</i>	<i>nxch6</i>	Number of 6-membered rings
<i>MDL187</i>	<i>Gmin</i>	Smallest atom E-State value in molecule
<i>MDL190</i>	<i>SHCsats</i>	Sum of hydrogen E-State on sp ³ C on saturated bond
<i>MDL210</i>	<i>SHBint2_Acnt</i>	Count of internal hydrogen bonds with 2 skeletal bonds between donor and acceptor

Table 5 Twelve Dragon descriptors employed in CAESAR model

Dragon Descriptor's code	Symbol	Definition
<i>DRA0107</i>	<i>PW5</i>	Path/walk 5—Randic shape index
<i>DRA0123</i>	<i>D/Dr06</i>	Distance/detour ring index of order 6
<i>DRA0341</i>	<i>MATS2p</i>	Moran autocorrelation—lag 2/weighted by atomic polarizabilities
<i>DRA0391</i>	<i>EEig10x</i>	Eigenvalue 10 from edge adj. matrix weighted by edge degrees
<i>DRA0451</i>	<i>ESpm11x</i>	Spectral moment 11 from edge adj. matrix weighted by edge degrees
<i>DRA0464</i>	<i>ESpm09d</i>	Spectral moment 09 from edge adj. matrix weighted by dipole moments
<i>DRA0551</i>	<i>GGI2</i>	Topological charge index of order 2
<i>DRA0565</i>	<i>JGI6</i>	Mean topological charge index of order6
<i>DRA0670</i>	<i>nRNNOx</i>	Number of N-nitroso groups (aliphatic)
<i>DRA0695</i>	<i>nPO4</i>	Number of phosphates/thiophosphates
<i>DRA0791</i>	<i>N-067</i>	Al2-NH
<i>DRA0802</i>	<i>N-078</i>	Ar-N = X/X-N = X

with skeletal branching pattern, both the electrotopological state (E-state) and molecular connectivity (Chi indices) structural descriptors were successfully implemented for prediction of genotoxicity and carcinogenicity (Contrera et al. 2003, 2005b). The authors (Votano et al. 2004) contend that one of the critical determining factors for good prediction results depend on nature of molecular structure representation employed in the model development process.

A complete set of *whole molecular descriptors* encode information on general structure features such as molecular size and shape, as well as specific information on skeletal variation and complexity. These structural features are expected to have a relationship to properties arising from intermolecular interactions and may also function to provide discrimination among multiple structural classes.

The atom-type, group-type, bond-type and single-atom E-state descriptors encode information on specific molecular features such as atom and bond types associated with important functional groups. Many of the descriptors relate directly to or associated with structural alerts as was reported in papers of Ashby and Tennant (1991) and Tennant and Zeiger (1993).

Some of E-state descriptors can be associated with structural alerts for carcinogenicity. For example, in Table 4 the *SdsN_acount* descriptor belongs to atom-type E-State account descriptors and expresses the count for the nitrogen atom type = N-associated with the azo group. The last one is also a structure alert and is correlated with carcinogenicity (Votano et al. 2004).

In Table 5 nRNNOx and N-078 descriptors are accounting for some specific fragments, whose presence is characterizing for the carcinogenic while the nPO4 descriptor accounts for non-carcinogenic class.

The global E-State descriptor Gmin is a measure of the most electrophilic atom in the molecule. Mechanistically, an electrophilic center is important for covalent bond formation with nucleophilic DNA. This is the reason why this descriptor was found between the most important descriptors correlated with carcinogenicity.

Hydrogen E-State descriptor SHCsats encodes E-state values for hydrogens on sp³ hybrid carbons bonded only with other sp³ carbon atoms. The electron accessibility of these sp³ hydrogens may relate in some manner to hydrophobic interactions between substrates and DNA or may have a relation to alkyl chlorides that are known toxicophores.

Thus, the descriptors used in our study refer to topological characteristics as well as to polarizability and charge distribution (related to reactivity).

Interestingly, some descriptors that we applied in our CAESAR models were also used by others authors (Contrera et al. 2005b; Votano et al. 2004) in carcinogenicity and genotoxicity modelling. It means that probably in future research it will be possible to find some common features for modelling carcinogenicity and genotoxicity.

It should be highlighted that the application of structure-information approach based on such descriptors like E-State has the following advantage: a model based on E-State descriptors (expressed as continuous value) can correlate carcinogenicity to a specific value of descriptor, whereas the use of fragment based structural alerts limits the model to a correlation of presence or absence of fragments or simple count of given fragments which can lead to false prediction for this reason.

The transparency of CP ANN algorithm using electro-topological MDL descriptors was demonstrated in the article by Fjodorova and Novič (2011).

A statistically-based method (counter propagation artificial neural network (CP ANN) was integrated with the knowledge-based one (structural alerts (SA) approach) to obtain the mechanistic interpretation of models. Mechanistic insight in CPANN models was demonstrated using the inherent mapping technique (i.e., Kohonen maps) which enables the visualization of the following features in 2D space: the carcinogenic potency; the distribution of descriptors in individual layers which express structural and electronic features such as molecular shape (linear, branched, cyclic, and polycyclic), bond length, taking into account electronic surroundings of molecules; and the distribution of congeneric groups of chemicals with indication of specific carcinogenic SAs with indication of broad mechanisms of action.

It was shown that some E-state descriptors relate directly or are associated with known SAs for carcinogenicity for such classes of chemicals like nitro compounds, nitro-aromatic, primary aromatic amines, and consequently carcinogens and non-carcinogens. Of course, not for all groups of chemicals clear clusters were obtained due to different mechanism of action inside one group of chemicals like in case of aliphatic halogens. But the advantage of the CPANN model is in a non-linear topological distribution of several small clusters of particular chemicals that are based on different modes of action.

The MDL topological, electrotopological and hydrogen bonding descriptors which express different aspects of shape and size of molecules encode information about electronic interactions of the atom and comprise features of electrostatic interaction between molecules. These important structural features contribute to the carcinogenicity and are expected to have a relationship to properties arising from intermolecular interactions and may also function to provide discrimination among multiple structural classes.

7 Conclusions

The implementation of REACH has provided an impetus to employ *in silico* models for the safety assessment of chemicals (EC 2008). Non-testing methods to assess genotoxic or carcinogenic hazard to humans include (Q)SARs as well as chemical grouping for read-across approaches.

Though read-across approaches are more frequently applied for cancer hazard assessment (ECHA 2014), (Q)SAR models represent the most formalized non-testing approach because it is grounded on the a large good quality database using robust scientific and statistical concepts. Read across in turn provide more specific information although it belongs to more subjective non-testing approach. A combination of (Q)SARs with read across was employed in the VEGA platform. The models such as VEGA, TOPKAT and MultiCASE give valid predictions of the presence and absence of genotoxicity/carcinogenicity (Jacobs et al. 2016). They are accepted for regulatory uses.

The models represented in VEGA (Fjodorova et al. 2010b) were built in accordance with 5 OECD principles for acceptance of QSAR models for regulatory use.

The CPDB rodent carcinogenic database was used for development of models for categorization of carcinogenic potency. Initial preprocessing of data and selection of data with carcinogenic potency for rats provided the consistent, homogeneous data suitable for QSAR modeling with carcinogenic potency response closer to human. The MDL and Dragon software programs were applied for calculating the molecular descriptors. The topological structure descriptors provided a sound bases for classifying molecular structures.

The CPANN algorithm was employed in modelling.

The statistical performance of models demonstrated good prediction statistics on the test set of 161 compounds with sensitivity of 75%, specificity of 61–69% in addition to accuracy 69–73%. A diverse external validation set of 738 compounds confirmed the robustness of our models regarding a large applicability domain, yielding the accuracy 60.0–61.4%, sensitivity 61.8–64.0%, and specificity 58.4–58.9%.

The new metric (Euclidian Distance (ED)) for evaluation applicability domain (AD) of neural network models was proposed.

A mechanistic interpretation of CPANN models was provided based on explanation of nature of descriptors used in the modelling with their possible chemical and/or biological activity. The integration of QSAR and SAR approach gave a solid fundament for robust prediction and mechanistic interpretation of obtained models.

The OECD has recently published new guidance principals for QSAR analysis of chemical carcinogens with mechanistic considerations (OECD 2015) for further assessment.

The carcinogenicity models incorporated in the VEGA platform can be used as a support in risk assessment, for instance, in setting priorities among chemicals for further testing.

VEGA models are described in detail in the help file of the software freely downloadable from the Vega website <http://www.vega-qsar.eu/>.

Acknowledgements The financial supports of the European Union through CAESAR project (SSPI-022674), the Slovenian Ministry of Higher Education, Science and Technology (grant P1-017) and the project LIFE PROSIL, LIFE12 ENV/IT/000154 are gratefully acknowledged.

References

- Ashby, J. (1985). Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity. *Environmental Mutagenesis*, 7, 919–921.
- Ashby, J., & Tennant, R. (1991). Definitive relationships among chemical structure, carcinogenicity and mutagenicity. *Mutation Research*, 257(3), 229–306.
- Bailey, A., Chanderbhan, N., Collazo-Braier, N., Cheeseman, M., & Twaroski, M. (2005). The use of structure-activity relationship analysis in the food contact notification program. *Regulatory Toxicology and Pharmacology*, 42, 225–235.
- Benfenati, E., Benigni, R., DeMarini, D., Helma, C., Kirkland, D., Martin, T., et al. (2009). Predictive models for carcinogenicity: Frameworks, state-of-the-art, and perspectives. *Journal of environmental science and health. Part C*, 27, 57–90.
- Benigni, R., & Bossa, C. (2008a). Structure alerts for carcinogenicity, and the Salmonella assay system: A novel insight through the chemical relational databases technology. *Mutation Research*, 659, 248–261.
- Benigni, R., Bossa, C., Jeliakova, N., Netzeva, T., & Worth, A. (2008b) The Benigni/Bossa rulebase for mutagenicity and carcinogenicity—a module of Toxtree. *EUR 23241 EN*. Retrieved July 18, 2016, from <http://ecb.jrc.ec.europa.eu/qsar/publications>.
- Benigni, R., & Bossa, C. (2008c). Predictivity of QSAR. *Journal of Chemical Information and Modeling*, 48, 971–980.
- Benigni, R., Bossa, C., & Worth, A. (2010). Structural analysis and predictive value of the rodent in vivo micronucleus assay results. *Mutagenesis*, 25(4), 335–341.
- Benigni, R., Bossa, C., Jeliakova, N., Netzeva, T., & Worth, A. (2008b). The Benigni/Bossarulebase for mutagenicity and carcinogenicity—a module of Toxtree. *European Commission report EUR 23241 EN*.
- Benigni, R., Bossa, C., Tcheremenskaia, O., & Worth, A. (2009) Development of structural alerts for the in vivo micronucleus assay in rodents. *EUR 23844 EN*. Retrieved July 18, 2016, from <http://ecb.jrc.ec.europa.eu/qsar/publications>.
- ChemFinder Ultra 10.0. (2009). CambridgeSoft Corp., Cambridge, MA.FDA, SAR Carcinogenicity database, Leadscope Inc., Columbus, OH.

- Combes, R., Grindon, C., Cronin, M., Roberts, D., & Garrod, J. (2008). Integrated decision-tree testing strategies for mutagenicity and carcinogenicity with respect to the requirements of the EU REACH legislation. *ATLA*, *36*, 43–63.
- Contrera, J., Hall, L., Kier, L., & MacLaughlin, P. (2005a). QSAR modeling of carcinogenic risk using discriminant analysis and topological molecular descriptors. *Current Drug Discovery Technologies*, *2*, 55–67.
- Contrera, J., Matthews, E., Kruhlak, N., & Benz, R. (2005b). In silico screening of chemicals for bacterial mutagenicity using electrotopological E-state indices and MDL QSAR software. *Regulatory Toxicology and Pharmacology*, *43*(3), 313–323.
- Contrera, J., Matthews, E., & Benz, R. (2003). Prediction the carcinogenic potential of pharmaceuticals in rodents using molecular structural similarity and E-state indices. *Regulatory Toxicology and Pharmacology*, *38*, 243–259.
- Cooper, J., Saracci, R., & Cole, P. (1979). Describing the validity of carcinogen screening test. *British Journal of Cancer*, *39*, 87–89.
- EC. (2008). Council Regulation (EC) No 440/2008 of 30 May 2008 laying down test methods pursuant to Regulation (EC) No 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH). *Official Journal L* *142*.
- ECHA. (2014). The use of alternatives to testing on animals for the REACH regulation. *Second report under Article 117(3) of the REACH Regulation*.
- Enslin, K., Gombar, V., & Blake, B. (1994). International commission for protection against environmental mutagens and carcinogens. Use of SAR in computer-assisted prediction of carcinogenicity and mutagenicity of chemicals by the TOPKAT program. *Mutation Research*, *205*, 47–61.
- Eriksson, L., Johansson, E., Wold, S., (1996). QSAR model validation. In: *Quantitative structure–activity relationships in environmental sciences VII. Proceedings of the 7th international workshop on QSAR in environmental sciences 24–28 June 1997* (pp. 381–397). Pensacola, FL, Denmark: SETAC Press.
- Eriksson, L., Jaworska, J., Worth, A., Cronin, M., McDowell, R., & Gramatica, P. (2003). Methods for reliability, uncertainty assessment, and applicability evaluations of classification and regression based QSARs. *Environmental Health Perspectives*, *111*, 1361–1375.
- Fjodorova, N., Vračko, M., Novič, M., Roncaglioni, A., Benfenati, E. (2010a). New public QSAR model for carcinogenicity. *Chemistry Central Journal*, *4*(1), 1–15. Retrieved July 18, 2016, from <http://www.journal.chemistrycentral.com/content/4/S1/S3>.
- Fjodorova, N., Vračko, M., Tuša, M., Jezierska, A., Novič, M., Kühne, R., et al. (2010b). Quantitative and qualitative models for carcinogenicity prediction for non-congeneric chemicals using CP ANN method for regulatory uses. *Molecular Diversity*, *14*(3), 581–594.
- Fjodorova, N., Vračko, M., Jezierska-Mazzarello, A., & Novič, M. (2010c). Counter propagation artificial neural networks categorical models for prediction of carcinogenicity for non-congeneric chemicals. *SAR and QSAR in Environmental Research*, *21*(1–2), 57–75.
- Fjodorova, N., & Novič, M. (2012). Integration of QSAR and SAR methods for the mechanistic interpretation of predictive models for carcinogenicity. *Computational and Structural Biotechnology Journal*, *1*(2).
- Fjodorova, N., & Novič, M. (2014). Comparison of criteria used to access carcinogenicity in CPANN QSAR models versus the knowledge-based expert system Toxtree. *SAR and QSAR in Environmental Research*, *25*(6), 423–441.
- Fjodorova, N., Novič, M., Roncaglioni, A., & Benfenati, E. (2011). Evaluating the applicability domain in the case of classification predictive models for carcinogenicity based on the counter propagation artificial neural network. *Journal of Computer-Aided Molecular Design*, *25*, 1147–1158.
- Fjodorova, N., & Novič, M. (2011). Some findings relevant to the mechanistic interpretation in the case of predictive models for carcinogenicity based on the counter propagation artificial neural network. *Journal of Computer-Aided Molecular Design*, *25*, 1159–1169.

- Fjodorova, N., & Novič, M. (2013). Rodent carcinogenicity dataset. *Dataset Papers in Medicine*, 1, 1–6.
- Golbraikh, A., & Tropsha, A. (2002). A: Beware of q²! *Journal of Molecular Graphics and Modelling*, 20, 269–276.
- Hall, L., & Hall, L. (2005). QSAR modeling based on structure-information for properties of interest in human health. *SAR and QSAR in Environmental Research*, 16(1–2), 13–41.
- Hall, L. (2004). A Structure-Information Approach to the Prediction of Biological Activities and Properties. *Chemistry & Biodiversity*, 1(1), 183–201.
- Helma, C. (2006). Lazy structure-activity relationships (lazar) for the prediction of rodent carcinogenicity and Salmonella mutagenicity. *Molecular Diversity*, 10(2), 147–158.
- IARC. (2016). *Monographs on the Evaluation of Carcinogenic Risks to Human*. Retrieved July 18, 2016, from <http://monographs.iarc.fr/ENG/Classification>.
- Jacobs, M., Colacci, A., Louekari, K., Luijten, M., Hakkert, B., Paparella, M., et al. (2016). International regulatory needs for development of an IATA for non-genotoxic carcinogenic chemical substances. *ALTEX*. Published online. Retrieved July 18, 2016, from http://www.altex.ch/resources/epub_Jacobs_of_1604272.pdf.
- Kazius, J., McGuire, R., & Bursi, R. (2005). Derivation and validation of toxicophores for mutagenicity prediction. *Journal of Medicinal Chemistry*, 48(1), 312–320.
- Kier, L., & Hall, L. (2005). The prediction of ADMET properties using structure information representations. *Chemistry & Biodiversity*, 2(11), 1428–1437.
- Kier, L., & Hall, L. (1999a). *Molecular structure description: the electrotopological state*. New York: Academic Press.
- Kier, L., & Hall, L. (1999b). The electrotopological state: structure modeling for QSAR and database analysis. In: J. Devillers & A. T. Balaban (Eds.), *Topological Indices and Related Descriptors in QSAR and QSPR* (pp. 491–562). Reading, UK: Gordon and Breach.
- Kier, L., & Hall, L. (2001). Database organization and searching with E-state indices. *SAR and QSAR in Environmental Research*, 12, 55–74.
- Lewis, D., Bird, M., & Jacobs, M. (2002). Human carcinogens: an evaluation study via the COMPACT and Hazard Expert procedures. *Human and Experimental Toxicology*, 21(3), 115–122.
- Maran, E., Novic, M., Barbieri, P., & Zupan, J. (2004). Application of counterpropagation artificial neural network for modelling properties of fish antibiotics. *SAR and QSAR in Environmental Research*, 15(5–6), 469–480.
- Marchant, C. (1996). Prediction of rodent carcinogenicity using the DEREK system for 30 chemicals currently being tested by the National Toxicology Program. The DEREK Collaborative Group. *Environmental Health Perspectives*, 104 (Suppl 5), 1065–1073.
- Matthews, E., & Contrera, J. (1998). A new highly specific method for predicting the carcinogenic potential of pharmaceuticals in rodents using enhanced MCASE QSAR-ES software. *Regulatory Toxicology and Pharmacology*, 28(3), 242–264.
- Matthews, E., Kruhlak, N., Cimino, M., Benz, R., & Contrera, J. (2006a). An analysis of genetic toxicity, reproductive and developmental toxicity, and carcinogenicity data: I. Identification of carcinogens using surrogate endpoints. *Regulatory Toxicology and Pharmacology*, 44, 83–96.
- Matthews, E., Kruhlak, N., Cimino, M., Benz, R., & Contrera, J. (2006b). An analysis of genetic toxicity, reproductive and developmental toxicity, and carcinogenicity data: II. Identification of genotoxicants, reproxicants, and carcinogens using in silico methods. *Regulatory Toxicology and Pharmacology*, 44(2), 97–110.
- Mazzatorta, P., Vračko, M., Jezierska, A., & Benfenati, E. (2003). Modeling toxicity by using supervised Kohonen neural networks. *Journal of Chemical Information and Computer Sciences*, 43, 485–492.
- Minovski, N., Župerl, Š., Drgan, V., & Novič, M. (2013). Assessment of applicability domain for multivariate counter-propagation artificial neural network predictive models by minimum euclidean distance space analysis: a case study. *Analytica Chimica Acta*, 759, 28–42.
- Netzeva, T., Worth, A., Aldenberg, T., Benigni, R., Cronin, M., Gramatica, P., et al. (2005). Current status of methods for defining the applicability domain of (quantitative) structure-

- activity relationships. The report and recommendations of ECVAM Workshop 52. *ATLA*, 33, 155–173.
- OECD. (2002). OECD *Environment, health and safety publications series on testing and assessment №35 guidance notes for analysis and evaluation of chronic toxicity and carcinogenicity studies*, Paris, France.
- OECD. (2004a). OECD principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationship models. Retrieved July 18, 2016, from <http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf>.
- OECD. (2004b). The report from the expert group on (quantitative) structure-activity relationships [(Q)SARS] on the principles for the validation of (Q)SARS. OECD SERIES ON TESTING AND ASSESSMENT, Number 49. ENV/JM/MONO(2004)24, 206. Retrieved July 18, 2016, from [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono\(2004\)24](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2004)24).
- OECD. (2007). Guidance document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models. ENV/JM/MONO(2007) 2 (pp. 154). Retrieved July 18, 2016, from [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono\(2007\)2](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2007)2).
- OECD. (2015). Fundamental and Guiding Principles for (Q)SAR Analysis Of Chemical Carcinogens With Mechanistic Considerations. Series on Testing and Assessment 229. Retrieved July 18, 2016, from [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono\(2015\)46&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2015)46&doclanguage=en).
- OECD TG 451. (2009). Carcinogenicity Studies. Retrieved July 18, 2016, from <http://www.oecdilibrary.org/docserver/download/9745101e.pdf?expires=1469696732&id=id&acname=guest&checksum=446DC8255BE3BC7EB69E834508462EEC>.
- OECD TG 453. (2009). Combined Chronic Toxicity/Carcinogenicity Studies. Retrieved July 18, 2016, from <http://www.oecd-ilibrary.org/docserver/download/9745301e.pdf?expires=1469696071&id=id&acname=guest&checksum=FE97C483DBD87757F34D3ADADECA817C>.
- Perkins, R., Rang, H., Tong, W., & Welsh, W. (2003). Quantitative structure– activity relationship methods: perspectives on drug discovery and toxicology. *Environmental Toxicology and Chemistry*, 22, 1666–1679.
- Peto, R., Pike, M., Bernstein, L., Gold, L., & Ames, B. (1984). The TD50: A proposed general convention for the numerical description of the carcinogenic potency of chemicals in chronic-exposure animal experiments. *Environmental Health Perspectives*, 58, 1–8.
- Prival, M. (2001). Evaluation of the TOPKAT system for predicting the carcinogenicity of chemicals. *Environmental and Molecular Mutagenesis*, 37, 55–69.
- Poroikov, V., Filimonov, D., & Lagunin, A. (2010). Multi-targeted natural products evaluation based on biological activity prediction with PASS. *Current Pharmaceutical Design*, 15, 1703–1717.
- Richard, A. (2004). DSSTox website launch: improving public access to databases for building structure-toxicity prediction models. *Preclinica*, 2(2), 103–108.
- Richard, A., & Williams, C. (2001). Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. *Mutation Research New Frontiers Issue*, 499(1), 27–52.
- Rose, K., & Hall, L. (2003). E-State Modeling of Fish Toxicity Independent of 3D Structure Information. *SAR and QSAR in Environmental Research*, 14, 113–129.
- Schüürmann, G., Kühne, R., Kleint, F., Ebert, R., Rothenbacher, C., Herth, P. (1997). A software system for automatic chemical property estimation from molecular structure. In F. Chen & G. Schüürmann (Eds.), *Quantitative Structure-Activity Relationships in Environmental Sciences*. VII (pp. 93–114). Pensacola, FL: SETAC Press.
- Schüürmann, G., Ebert, R., Nendza, M., Dearden, J., Paschke, A., Kühne, R. (2007). Prediction of fate-related compound properties. In K. van Leeuwen & T. Vermeire (Eds.), *Risk Assessment of Chemicals. An Introduction* (pp. 375–426). Dordrecht, NL: Springer Science.
- Serafimova, R., Todorov, M., Pavlov, T., Kotov, S., Jacob, E., Aptul, A. A., et al. (2007). Identification of the structural requirements for mutagenicity, by incorporating molecular

- flexibility and metabolic activation of chemicals. II. General ames mutagenicity model. *Chemical Research in Toxicology*, 20(4), 662–676.
- Serafimova, R., Gatnik, M., Worth, A. (2010). Review of QSAR models and software tools for predicting genotoxicity and carcinogenicity, *JRC 59068. EUR 24427*, 49. Retrieved July 18, 2016, from https://eurl-ecvam.jrc.ec.europa.eu/laboratories-research/predictive_toxicology/doc/EUR_24427_EN.pdf.
- Tennant, R., & Zeiger, E. (1993). Genetic toxicology: current status of methods of carcinogen identification. *Environmental Health Perspectives*, 100, 307–315.
- Valerio, L., Arvidson, K., Chanderbhan, R., & Contrera, J. (2007). Prediction of rodent carcinogenic potential of naturally occurring chemicals in the human diet using high-throughput QSAR predictive modeling. *Toxicology and Applied Pharmacology*, 222(1), 1–16.
- Votano, J., Parham, M., Hall, L., Kier, L., Orloff, S., Tropsha, A., et al. (2004). Three new consensus QSAR models for the prediction of ames genotoxicity. *Mutagenesis*, 19, 365–378.
- Vračko, M., Mills, D., & Basak, S. (2004). Structure-mutagenicity modeling using counter propagation neural networks. *Environmental Toxicology and Pharmacology*, 16, 25–36.
- Vračko, M., Novič, M., & Zupan, J. (1999). Study of structure-toxicity relationship by a counterpropagation neural network. *Analytica Chimica Acta*, 384(3), 319–332.
- Woo, Y., Lai, D. (2005). OncoLogic: A mechanism-based expert system for predicting the carcinogenic potential of chemicals. In C. Helma (Ed.), *Predictive Toxicology* (pp. 385–413). Boca Raton FL, USA: CRC Press.
- Zupan, J., & Gasteiger, J. (1999). *Neural networks in chemistry and drug design* (2nd ed.). Weinheim: Wiley-VCH Verlag GmbH.
- Zupan, J., Novic, M., & Ruisainchez, I. (1997). Kohonen and counterpropagation artificial neural networks in analytical chemistry. *Chemometrics and Intelligent Laboratory Systems.Tutorial*, 38, 1–23.

Big Data in Structure-Property Studies—From Definitions to Models

Jaroslav Polanski

Abstract What is big data and how important is big data in drug design? We analyze the big data types that are available in drug design as well as the methods that are used for their analyses. In particular, we discuss the definitions of the substantial molecular data concepts of a property and a descriptor to distinguish molecular big data types. The fact that measured property data are seldom available often requires property predictions. At the same time, this *property deficit* is among the main obstacles that limit big data structure-property studies.

Keywords Big data • Business intelligence • Chemical databases • Chemical space • CoMFA • Data analysis • Data binning • Data management • Data populations in chemistry • Economics • Google algorithm for spread of flu prediction • Hierarchy of scientific explanation • logP • Molecular big data architectures • Molecular big data by a large number of objects with a single property annotation (PE-LPA) • Molecular big data by descriptor expansion (DE) • Molecular big data by property expansion (PE) • Molecular big data by the increase of objects with predicted property annotation (PE-PPA) • Molecular descriptor • PASS • Property • Property deficit • QSAR • QSPR • Quantitative structure-economics relationship

1 Introduction

Chemistry attempts to find the rules that control the behavior of chemical compounds. Preferably, for the universal laws, e.g., conservation energy law, this refers to a whole population of molecules and/or substances. The relationship between the structure and a property of a chemical compound is an essential concept in

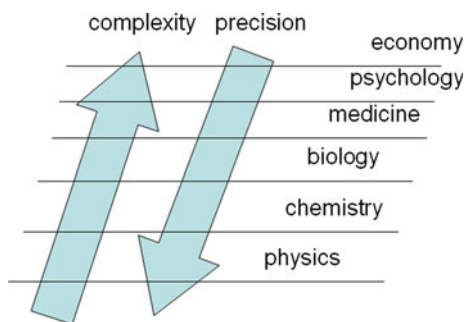
J. Polanski (✉)
Institute of Chemistry, University of Silesia, 9 Szkolna Street,
40-006 Katowice, Poland
e-mail: jaroslav.polanski@us.edu.pl

chemistry that guides, for example, drug design. This problem is addressed by classical QSAR (QSPR) which, however, only describes a small series of congeneric compounds, usually chemotypes, i.e. structurally related molecules. Therefore, in practice we usually analyze chemical data that cannot give us a broad-spectrum structure-property mapping, which results in unconvincing and weakly innovative projects. With an enlargement of the chemical space, we could modify the questions that are asked. For example, we were interested in whether a general rule exists that would differentiate drugs from non-drugs and which molecular descriptors determine this. Can anything like drug-like molecules be identified and what does drug-likeness mean? At the same time, the availability of computers has resulted in an explosion of information. Accordingly, we realized that data has also become bigger and bigger in chemistry.

There are many definitions of big data but generally what determines the difference between conventional and big datasets are *volume*, *velocity* and *variety*, where volume refers to the massive size of datasets; velocity to the rate of the increase of information and variety to the diverse data forms. Alternatively, big data is sometimes defined by a high degree of information complexity (Laney 2016), which causes traditional methods to fail when they are used for processing.

How should big data be gathered and managed? What questions should be asked in order to address and answer actual problems? In science, we often need reductionism to answer questions. This implies that a certain field of study relies on the disciplines that focus on less complex systems. Thus, chemistry is founded on physics, while biology needs chemistry in order to understand its molecular mechanisms. This is defined as the so-called hierarchy of scientific explanation. In this concept, scientists use a reasoning mode by using the tools of the underlying science. Therefore, in order to explain psychology, we need biology and medicine, which in turn can be explained by chemistry, which is situated over physics, which relies on the empirical facts and is located at the very bottom (Rosenblum and Kuttner 2011). The newest discipline of *behavioural economics* investigates the *effects of psychological, social, cognitive and emotional factors on economics* (Kahneman 2011). These objectives include a variety of topics involving, for

Fig. 1 The increase in information complexity (precision decrease) from physics to economy



example, psychology; therefore; situating economics on the top position in this hierarchy. Accordingly, complexity increases here from physics to economics, while precision of the model moves in backward direction (Fig. 1). Therefore, if we try to find the most complex and challenging applications of big data, economics is at the top. However, at the same time, this area will also provide us with a manifold and multi-dimensional insight into a variety of problems. For example, just to illustrate the importance of economics in drug design, we should realize that we need economic considerations to fully understand the fate of drugs on the market. In fact, only this understanding can make decision-making in pharma effective enough. Accordingly, the first quantitative structure-economics relationship that is based on big data and merges chemical descriptors and economic indicators has just been explored (Polanski et al. 2016a).

Business intelligence is a term that is sometimes used to describe the variety of IT technologies that are used to understand, explain and make decisions in economics. From the economics and business point of view, drug design is a part of medicine and healthcare, which is one of the most advanced sectors in the current economy, where the so-called *evidence-based medicine* is a hard data based trend here that can contribute to (Maheshwari 2014)

- Patient diagnostics
- Treatment evaluation
- Wellness management
- Fraud and abuses
- Public health management

Can analyses of big data significantly improve our wellness in the context of drug design? A good example that can provide a positive answer here is thoroughly discussed in reference (Cukier and Mayer-Schönberger 2013). Google, which processes as much as 24 petabytes (peta means 10^{15}) of information per day, is probably the largest available data store that is currently available. By analyzing what people queried on the Internet, Google was able to efficiently predict the spread of flu (Ginsberg et al. 2009). It was the pattern of the big data describing the search history that enabled this analysis. It is worth mentioning here that the original Google algorithm was questioned, which however brought about its improvement (Lazer et al. 2014).

The role of big data and advanced analytics in drug discovery, development and commercialization has recently been carefully analyzed by Slezák et al. (2014), who indicated several big data sources in health care, i.e.

- Claims
- Clinical
- Sales and dispensing
- Clinical research, safety and pharmacovigilance
- Patient generated data

It is generally believed that big data will bring new value and innovation. For example, Szlezák et al. cited the recent McKinsey research that suggests that the potential use of the big data in the US health care could reduce costs by \$300 billion a year. However, this kind of information is also much less clearly defined and messy. Accordingly, its analysis causes serious problems. The first of which is that conventional statistics is designed for conventional data. We are not aware enough of the differences and we are not ready to change the addiction to small data sets. The analysis of big systems allows us to observe the details we would never have detected using traditional small data sampling. However, measurement errors can be much larger for big data. It is hard to believe that the accuracy or exactitude can be similar to that of conventional traditional (*small*) data processing. This also means that there is an enormous complication in explaining, modeling or understanding the medical, pharmacological or chemical effects represented by big data. Bio- and chemoinformatics are tools that are used for in silico data processing in chemistry, pharmacy and medicine. Chemoinformatics was originally defined as the combination of all of the information resources that a scientist needs in order to optimize the properties of a ligand in order for it to become a drug. Similar goals are targeted by bioinformatics but the focus of the latter discipline is more biologically oriented. Therefore, bio- and chemoinformatics design tools for drug discovery and development. More recently, what can be observed is that both terms are merging into a single discipline. This discipline should also coordinate the search for the solutions for managing and processing big data in drug design.

Accordingly, what precisely does the term (big) data mean? What is the difference between traditional and big data systems? Last but not least, where can we find this in chemistry and chemo- or bioinformatics? The broadest definition interprets data as *anything that is recorded*. This also involves *metadata*, i.e., data that refers to other data. This means that data can be both *ordered* and *unordered* collections of values. Moreover, the values can be both *nominal* and *numerical* values, whereas the latter can be discrete numbers, intervals or ratios. Another type of data (Binary Large Objects, BLOBs) is used to describe audio, video and graphic files. A special analysis type is used to explore these (Maheshwari 2014). In drug design, another data feature is of crucial importance. This is the data ownership attribute, which can cause a problem with data availability. Although we are aware of how important it is to share data with limited or unlimited parties (sharable data; data sharing problem), it is still much more common that *data in confidence* is the reality in pharma R&D. The decreasing efficiency in pharma has inspired the collaborative drug design projects where all of the collaborators that are involved share the data among themselves (CDD, collaboratedrug.com). Although data sharing between traditional pharma companies is still a matter of controversy, more and more people are convinced that this could significantly improve the efficiency in this field. Let us now try to identify and define big data sources as well as the methods that are used to manipulate big data in chemoinformatics.

2 Molecular Definitions and Data Populations in Chemistry and Drug Design

Datafication is a term that drew our attention to the fact of the growing importance of data acquisition and management. We can also observe this trend in chemistry. Accordingly, the term chemical space (CS) has recently appeared. In its broadest sense, the CS can be interpreted as a structure to organize chemical data. As the majority of data in chemistry can be connected directly to chemical compounds and molecules, *molecular data* is a substantial type of information and chemists have realized the need to bring some organization into the molecular world with the increasing importance of *molecular design*. Accordingly, the chemical space (CS) is first of all a concept that is designed to organize a whole population of chemical compounds. In the context of big data, we should realize that the whole molecular population is included here. The original definition was rather vague and drew our attention to a cosmological analogy; therefore, the universe space is populated by stars, where the chemical space is populated by chemical compounds. Accordingly, Lipinski and Hopkins (2004) coined the phrase that chemists are *navigating chemical space*. Actually, we should see CS as a structure for the mapping of chemical compounds by molecular data, i.e., descriptors or indicators relating to molecular structures and properties that can be measured in experiments (Polanski 2009a; Polanski and Gasteiger 2016).

An attempt to define the CS that is related to mathematics is illustrated in Fig. 2. Thus, the CS is a structure for the arrangement of chemical compounds (Polanski 2009a). Since the term ‘chemical compounds’ includes both molecules and substances, both representations are recorded in the CS. The CS can be divided into two basic moieties—the factual chemical space (FCS) that is described by the chemical compounds that have already been obtained and registered and the virtual CS (VCS) that maps potential substances. The CS structure can be used to design

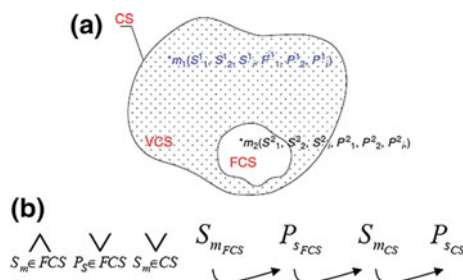


Fig. 2 Chemical space (CS) formed by putting together factual (FCS) and virtual (VCS) chemical compound data represented by descriptors S and properties P (a). An operator shown below provides formal notation for QSAR, therefore, in QSAR domain for the substances in FCS we are mapping structure (S) to property (P) to further use the modeled relationship to find new structures (S) now in CS of the designed properties (P). Modified from Polanski (2009a) and Polanski and Gasteiger (2016)

mathematically inspired operators to illustrate and explain chemical problems. Accordingly, we use this notation to define QSAR (QSPR) modeling where we used a series of FCS compounds with the measured property (biological activity) to model the QSAR function, then we used this function to predict the activity of *novel compounds* (which in fact can be both FCS or VCS elements) that should, of course, be kept within the QSAR domain. Both of these steps were performed in silico (Fig. 2b). We can use this concept to illustrate the chemical operators that occur in vitro, e.g., chemical synthesis, or both in vitro and in silico, e.g. drug design or synthesis design problems.

Because mathematics, including mathematics in chemistry, needs precise definitions and a deep understanding of the molecular concepts, we will now focus on this. Chemical compounds, or more precisely, chemical compound data can be represented in the CS by descriptors or properties, which are two possible chemical record types. The term property in chemistry is precisely defined by IUPAC. Therefore, we understand a property to be

A set of data elements (system, component, kind-of-property) common to a set of particular properties, e.g. substance concentration of glucose in blood plasma. Information about identification, time and result is not considered.

Properties are *measured* and such an operation is also a precisely IUPAC defined operation.

A description of a property of a system by means of a set of specified rules, that maps the property onto a scale of specified values, by direct or 'mathematical' comparison with specified reference(s). The demand for rules makes 'measurement' a scientific concept in contrast to the mere colloquial sense of 'description'. However, in the present definition, 'measurement' has a wider meaning than given in elementary physics. Even a very incomplete description of, for instance, a patient (at a stated time) has to be given by a set of measurements, that are easier to manage and grasp.

The term *molecular descriptor* has been defined by Todeschini (definition *DES*) to describe all of the numbers that relate to molecules that are obtained as:

A final result of a logic and mathematical procedure transforming chemical information encoded within a symbolic representation of a molecule into a useful number (Todeschini and Consonni 2000).

This gives us a chance to distinguish between a molecule and a substance as two counterparts of the description of a chemical compound that differentiate between *experimental measurements*, i.e., *properties* (mainly for real substances) and other numbers that describe in silico *simulated* molecules or substances, namely, molecular descriptors. However, Todeschini is not consistent here and we can read, for example:

[...] molecular descriptors are divided into two main classes: experimental measurements, such as logP, molar refractivity, dipole moment, polarizability, and, in general, physico-chemical properties and theoretical molecular descriptors (Consonni and Todeschini 2010).

Interestingly, the latter definition combines properties and descriptors into a single class of *molecular descriptors*.

In reverse, if we compare the IUPAC definition, it does not differentiate between descriptors and properties and defines the *structure-property correlations* as follows:

Structure-property correlations (SPC) refer to all statistical mathematical methods used to correlate any molecular property (intrinsic, chemical or biological) to any other property, using statistical regression or pattern recognition techniques.

Accordingly, in the *SPC* definition, the single category of *property* represents both descriptors (according to the *DES* Todeschini definition) and properties.

For the requirements of (Q)SAR and (Q)SPR, it is, however, highly recommended to clearly distinguish between descriptors and properties. In the context of the IUPAC definition, these categories are more or less defined as *properties referring to a molecule (intrinsic property or molecular descriptor) or a substance (chemical or biological property)*, respectively. For a broader discussion of the problem, including the complications within the identification of the meaning of chemical compounds, molecules and substances in chemistry, the reader can compare (Polanski and Gasteiger 2016). Therein, the reader can also find the examples that illustrate how precisely differentiating properties from descriptors can be in chemistry. For example, let us ask a question about *molecular weight* (MW). Is it a property or a descriptor? According to our definitions, it can be both a property and descriptor. Therefore, if measured in the experiment, e.g., in a MS spectrometer *in vitro*, this is a property. However, we usually calculate MW simply by summing the MWs of the atoms that form a molecule, which means it is molecular descriptor of the dimension *zero* (Polanski 2009a; Polanski and Gasteiger 2016). Both numbers will be *practically*, but not *theoretically*, the same.

The differentiation of properties versus descriptors is of substantial importance for understanding molecular design where we can make *property predictions* on the basis of the calculations of *molecular descriptors* for *hypothetical molecular structures* (*in silico*) which are not available for *in vitro* substance measurements. It should further be realized that in reality there is not much difference when we are designing in the FCS or VCS sub-spaces, because FCS compounds are not usually available for measurements *in vitro*. They have been obtained and registered in databases or literature but are not usually available any longer. The low availability of chemical compounds in the FCS is a major problem in molecular design. Moreover, despite the common belief, property measurements in chemistry are rare (Polanski and Gasteiger 2016). The reason for this is economics. Measurements are expensive because they need not only the measurement step itself but also the synthesis and often purification of chemical compounds. These problems were realized and large chemical compound libraries have recently been offered commercially on the market to close this gap.

Let us now ask the question how big are the molecular populations of the chemical compounds, in which chemical compounds represent both molecules and substances? Figure 2a is an illustration of the chemical space consisting of its

factual and virtual parts. The common belief is that a large molecular population has been described in organic chemistry. In fact, the Chemical Abstract Service (CAS) registers slightly more than 100 mln compounds, i.e., ca. 10^8 . *Of the 100 million substances in the CAS REGISTRY, approximately 75 million were added over the past ten years. On average, CAS has registered one substance every 2.5 min over the past 50 years* (CAS). For a better illustration, let us compare this to the human population, which has reached ca. 7.4 billion, i.e. ca. 10^9 . How does the FCS population compare to the virtual chemical space fraction VCS? Different numbers are cited for the latter that range from 10^{60} to 10^{200} . This means only a small fraction of the potential molecular population has currently been described.

On the other hand, Wang et al. (2014) estimated the data connected to the measured property values to:

- 700,000 bioassays
- 200,000,000 bioactivity outcomes
- 1,200,000,000 data points
- 2,800,000 small molecule samples
- 1,900,000 chemical structures
- 108,000 RNAi reagents

Compare also Cheng et al. (2014) for the bibliometric analysis of the PubChem database. Properties are stored in chemical databases. Basically, a good example of a property database is the CAS, which registers the descriptors and properties for real compounds. In particular, the CAS registers all of the chemical compounds that have ever been reported in the literature, i.e., the VCS part of CS. If we would like to include a structure of the molecular data offered by the CAS, we can see that this is a typical big data where the number of analyzed objects (chemical compounds) determines the volume of information.

The molecular databases that are available are listed in Table 1. Both descriptors and properties are registered here, however, as could be expected molecular properties are the most essential records for the FCS compounds. However, molecular descriptors are in fact also an important information type here, e.g., compare the 3D molecular structures predicted (simulated in silico) to the molecular docking studies in the ZINC database.

The properties that are actually measured are by definition not available for VCS compounds. We should remember that they are also rare for most of FCS population. Thus, we often have to replace them with predicted values. In this context, the CAS service has recently also offered the predicted property values for chemical substances. The problem of the big data that is generated by predicted properties will be addressed in detail in Sect. 4.4.

Besides the data discussed above, the molecular populations in chemistry and drug design can be arranged into another type of big data. These are the data where a relatively low number of objects are described by an extremely large number of molecular descriptors. The best examples of such data are descriptors generated by the Molecular Interaction Filed (MIF) method for 3D QSAR studies, e.g.,

Table 1 Examples of chemical databases recording molecular big data

Database	Molecular population	Descriptors	Properties	Remarks
Chemical Abstracts Service: CAS REGISTRY CAplus CASREACT CHEMCATS CHEMLIST CIN MARPAT https://www.cas.org/content/casdatabases	117 M small molecule data (CAS REGISTRY)	Coding descriptors, chemical names, Markush structures (MARPAT)	Chemical structures, predicted and measured properties, tags, spectra	The broadest source harvested chemical data including journals, patent authorities, web sources; chemical literature from 1800; mainly FCS or if predicted (V)CS data
Reaxys http://www.reaxys.com	Merged databases 9 M (<i>Beilstein</i>), 2.5 M (<i>Gmelin</i>) and <i>Patentis chemistry</i>	Coding descriptors, chemical structures and substructures	Chemical and physical properties, preparative methods, chemical behavior, predicted chemical reactivity (synthesis design)	A searchable database of some 10 million reactions and properties; chemical literature from 1771; basically FCS data
ZINC http://zinc.docking.org	Over 2.7 M commercially-available compounds ca. 90 M <i>compounds that can simply be purchased</i>	Coding descriptors, 3-D structures (simulated in silico)	3-D structures (measured); activity and pIC50 data, if available	Free database for virtual screening in ready-to-dock, 3D format; subsets by vendor and other criteria such as Lipinski-compliant (Lipinski 2000), lead-like (Schneider 2002; Teague et al. 1999) and fragment-like (Verdonk et al. 2003) compounds; (V)CS and FCS compounds
ChEMBL European Bioinformatics Institute http://www.ebi.ac.uk/chembl/	Targets: 11,019; omponents records: ca. 2 M; distinct compounds: 1,592,191; activity data 13,967,816	Coding descriptors; chemical and pharmacological names	Activity and pIC50 data	FCS compounds
eMolecules http://www.emolecules.com	Over 8.0 M unique substances	Coding descriptors, structures and substructures; chemical and pharmaceutical names	Supplier and price data (chemical like catalogue)	FCS compounds commercially available

(continued)

Table 1 (continued)

Database	Molecular population	Descriptors	Properties	Remarks
PubChem http://pubchem.ncbi.nlm.nih.gov	Substances 140 M substance tested 3 M bioactivity data 226 M bioassay 1 M protein target ca. 6 (Cheng et al. 2014)	Coding descriptors, structures and substructures; chemical and pharmaceutical names	Bioactivity data chemical properties	FCS compounds compounds, bioassay, bioactivity, target, patent data
GDB Databases http://gdb.unibe.ch/downloads/	Over 977 M molecules	Coding descriptors		VCS databases: GDB-11 enumerates small organic molecules up to 11 atoms of C, N, O and F following simple chemical stability and synthetic feasibility rules. GDB-13 enumerates small organic molecules up to 13 atoms of C, N, O, S and Cl following simple chemical stability and synthetic feasibility rules. With 977 468 314 structures, GDB-13 is the largest publicly available small organic molecule database to date (Blum and Reymond 2009)

Comparable Molecular Field Analysis (CoMFA). In such a study, a single molecule is represented by a series of calculated MIF descriptors whose numbers can reach as much as an order of 10^3 or even 10^4 . Therefore, a typical CoMFA study engages no more than 10^2 molecules described by a much larger descriptor data series. These data types are big based on their complexity and not the number of objects engaged. It is worth mentioning that since all descriptors are calculated here and 3D molecular representations are also simulated *in silico*, the majority of data are not connected with the FCS data. The most important data of this type are generated by receptor independent or receptor dependent 3D-7D QSAR (Polanski 2009b), molecular dynamics, molecular docking, etc. Typically, these data are generated and processed *in silico*, but are not stored for further use and/or control. The big data that is generated by molecular descriptors will be analyzed in Sect. 4.1.

Accordingly, we can now attempt to categorize the data types that we can encounter as big records. Let us define the data as simply the collection of information that is formed by records. As was discussed above, this can grow big through an increase in the number of objects, by an increase in the number of variable entries that describe an individual object or by an increase in both the objects and observables (Fig. 3). If we now focus our attention on chemistry, we can further observe that several basic data variants can be formally indicated. This involves:

- properties measured for FCS substances
- properties predicted for FCS or VCS substances
- descriptors calculated for FCS or VCS molecules

Alternatively, we can more precisely indicate systems in which a big size is determined by

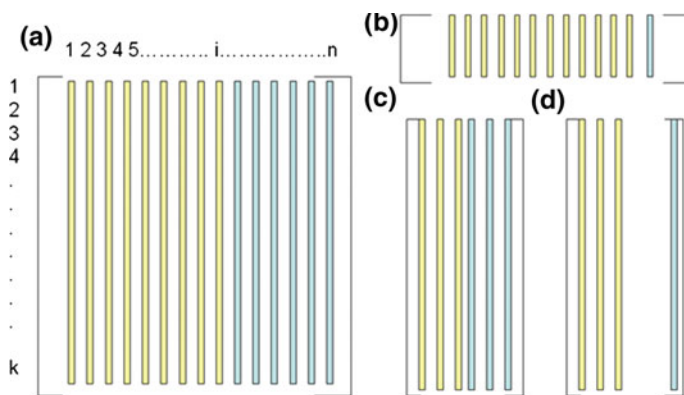


Fig. 3 A scheme of data architecture in drug design. Descriptors (*yellow*) and properties (*blue*) are arranged horizontally (1 to n) for the objects put together vertically (1 to k) (a). In classical (multidimensional) m-QSAR the data are usually dominated by descriptors, *descriptor expansion* (b), ideal big data composed for a large number of objects annotated by the equable numbers of descriptors and properties (c), actual situation where the data get bigger by the increase of property annotation in a single property column (d)

- Property annotation expansion
- Predicted property annotation expansion
- Descriptor expansion

For better clarity, we will code these categories as PE (property expansion) or DE (descriptor expansion). As we explained before, measured properties are rare, therefore, the structure of PE is usually large due to the number of objects that are annotated with a single property type and not by a large number of property types. Conversely, the variety of descriptors that are designed to describe molecules makes DE data large due to the number of descriptor variables. Moreover, *the deficiency of properties* means that the measured property values must often be replaced by the predicted property values. This creates a specific data type that we will refer to as the PPA (predicted property annotation) data, which will be discussed in detail in Sect. 4.4.

3 From Molecular Big Data to Knowledge

Processing big data requires efficient *data management*, e.g., searches or screening, etc. Therefore, molecular databases are indispensable. It is worth mentioning that molecular descriptors, in particular, molecular and chemical formulae often replaced recently with SMILES codes, are generally used in databases for addressing the individual chemical compounds in FCS. What about VCS? This part is also interesting as a potential source of novel chemical compounds. Accordingly, the generation of virtual drug candidate libraries is an important problem (Blum and Reymond 2009; Reymond 2015). Therefore, the databases of virtual molecules are as important as those of real compounds. The main difference between FCS and (V)CS chemical databases is the fact that the latter cannot register measured properties. Thus, the most important information here is the molecular address, i.e., the molecular descriptors coding this individual molecule. In fact, while constructing VCS databases we usually do not care if all elements of the database have never been registered among the FCS structures, which we indicated here by the notation (V)CS. Molecular (V)CS databases do not differ from those that register real molecular bodies with properties in their information structure, i.e., they register a large number of molecular objects, i.e., millions 10^6 or billions 10^9 . Compare Table 1 for the examples of the individual databases.

Data analysis is the next problem. Statistics developed a variety of methods that allowed us to explore data and convert them into knowledge. Formally, this can be grouped into several categories of which those enumerated below are often used in chemistry, chemoinformatics and chemometrics:

- Basic statistics, e.g. mean, sum, regression, correlation analysis
- Neural networks
- Pattern recognition

- Dimensionality reduction (feature selection, feature reduction)
- Principal component analysis (PCA)
- Partial least squares (PLS)
- Fourier analysis
- Wavelet transformations
- Machine learning
- Experimental design

For further analysis the reader can compare reference (Gasteiger 2003; Gasteiger and Engel 2003; Polanski and Gasteiger 2016).

Among these, let us analyze a single regression to show how important and informative simple statistics can be for data processing. Regression is a tool that can be used not only for data modeling but also for data compression. As the compression capability may not be understandable, let us focus on the example of experimental data that is recorded by two variables noted by a matrix of the individual parameters of size $10,000 \times 2$, which results in 20,000 numbers. This data if modeled using regression can replace this by two numbers (*regression*) or even a single number for *ridge regression* (Polanski 2009a; Polanski and Gasteiger 2016). This method is also the most important method for predicting and forecasting, i.e., design, thus also illustrating the potential power of regression for big data analysis. However, simple regression as a modeling method is often insufficiently robust and is prone to errors and therefore too *fragile* to process big data.

The methods enumerated below are usually associated with processing big data (Ldtopology; Polanski et al. 2016a).

- Projection
- Feature Selection or Extraction
- Clustering
- Classification, in particular data binning

In *projection* the original high dimensionality vector space is transformed to align in a novel lower dimensionality space that can provide patterns that are easier to interpret. At the same time, the original relationships between the observables during projection can be distorted but the extent of the distortion should be as low as possible. Principal component analysis (PCA) and self organizing maps (SOM) are typical examples of this method. Formally, we can also divide projections into linear and nonlinear ones.

Feature selection or extraction focus on the elimination of the part of high dimensionality data that cannot be correlated with the analyzed target signal. In other words it is designed to eliminate a *noise*. This method is often used as preprocessing for other algorithms. Last but not the least, it can also be used to reduce the dimensionality of the signals because with an increase in the data size, the models become more precise (internal data modeling) but at the same time much less predictive (new data). This problem is often referred to as *over-fitting*. Practically, the reduction of data dimensions also makes data processing less expensive.

The term *classification* is used to define a variety of algorithms which results in labeling the data in such a way that *reasonable* relations can be discovered in the analyzed data. A variety of individual methods have been developed by the so-called machine learning techniques generally based on *supervised learning* where the latter means that a computer is instructed (trained) what is expected from *reasonable labeling* during the process.

Clustering differs from classification in that data are distributed into output groups (*clusters*) without supervising (*unsupervised learning*), which more or less means that the data are not labeled. Once more, a variety of individual methods have been developed for data clustering.

Data binning is sometimes used as a synonym of classification but will be here used in the more specific way to refer to the preprocessing method. How are data treated when there are a large number of measurements available? For example, temperature measurements around the globe are analyzed to evaluate the importance of the alleged effect of global warming. These analyses are often based on the mean values of temperatures that are averaged for a certain month through several years. Such a procedure is called data binning. The formal definition reads as follows: binning or bucketing is a method in which the *original* [continuous] *data values which fall into a given small interval, a bin, is replaced by a value representative of that interval*. Accordingly, we often use monthly binned data for the assessment of global temperature changes. Actually, data binning is typically used to understand and explain the substantial effects on our everyday life.

4 Molecular Big Data Architectures

Figure 3 presents the possible architectures of molecular data schematically. Therefore, the chemical descriptors S and properties P of the chemical compounds probed in the chemical space (Fig. 2) are merged to form the data matrix in Fig. 3a. A special type of molecular descriptors is used to code the molecules, e.g., SMILES. Theoretically, we can have several situations depending on the number of properties and descriptors that are registered in the matrix in Fig. 3. Thus, the matrix can be balanced when the number of descriptors and properties is similar (Fig. 3a). Alternatively, the descriptors or properties could dominate in unbalanced cases.

Further, the data matrix can get bigger by horizontal (Fig. 3b) or vertical (Fig. 3c, d) expansions. In practice, however, molecular data are always unbalanced with the domination of molecular descriptors horizontally and usually a single property column is available in both cases (Fig. 3b, d).

The vertical data matrix expansion (Fig. 3d) resulting from an increase in the number of molecular objects increases at the same time the number of the single property annotations that have been measured.

4.1 *Molecular Big Data Resulted from the Descriptor Expansion (DE)*

The DE data architectures are the most common examples of big data in drug design. Descriptors are relatively easily available from calculations or computer simulations and are clearly much cheaper than experimental measurements. This means that we can meet such data types very often.

Typical applications are enumerated below:

- 3D-7D QSAR
- Molecular interaction Mi-QSAR
- Genomic (structure-target) QSAR
- Microarray (genomic) data analyses
- Structure-target (multiple target) QSAR
- Docking analyses, e.g. Comparative Molecular Binding Analysis COMBINE

The data processing methods used for these analyses are relatively well recognized. Typical data usually involves fewer than 100 molecules for a single property, typically biological activity value, and a large number of descriptors. We are aware that in drug design the credibility of the analysis increases with the number of molecules that are probed in an analysis and decreases with an increase in the number of descriptors involved.

Partial Least Squares data modeling with data projection to latent variables are the typical statistics that are used. The crucial steps involve

- splitting data into the so-called training and test sets
- PLS modeling
- optional feature selection (data reduction step)
- evaluation of the predictive power of the model
- property prediction for an external set of novel compounds
- optional synthesis of novel compounds
- optional comparison of the predicted versus measured property values for newly synthesized compounds (test set)

In comparison to 3D QSAR in which conformations remain stable in time, 4D QSAR analyses increases in the number of descriptors by the molecular dynamics conformations (*poses*) that are generated. Further, QSAR dimensions also increase the number of data involved.

Illustrative examples of property prediction in Comparative Molecular Field Analysis (CoMFA) and Comparative Molecular Surface Analysis (CoMSA) modeling are shown in Fig. 4. The data reduction step in CoMSA versus CoMFA can significantly improve the robustness of the models resulted.

The most important question that appears in DE studies is whether the descriptors that are calculated for the molecular objects are really independent values. In fact, calculation often means that the mathematical function that connects the calculated values to some original ones is known. We will not discuss the

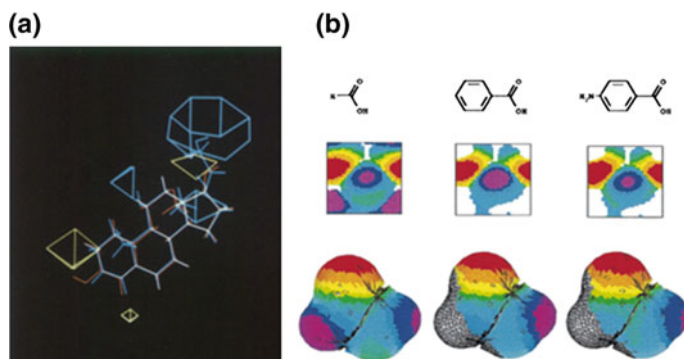


Fig. 4 Visualization of m-QSAR molecular prediction for the descriptor expanded big data in the CoMSA and CoMFA analyses. Modified from Cramer et al. (1988) and Polanski and Walczak (2000)

problems of the DE type data further and the reader should go to the widely available references in this field.

4.2 Molecular Big Data Resulting from Property Expansion (PE)

Alternatively to the DE scheme data could get big through the expansion of the number of properties that are measured.

The PASS method, Prediction of Activity Spectra for Substances (PASS) approach is a method that declares the importance of the analyses of the whole activity spectra of chemical compound, i.e., a compound's interaction with various biological systems. However, in reality a single property is usually targeted and examples in which more property values are analyzed can be only found occasionally. For example, we indicated a *property deficit* relatively early and decided to investigate the biological activity spectrum instead of a single activity in experimental practice, e.g. for quinolines (Musiol et al. 2007).

At the same time, PASS is designed for property prediction rather than for registering the actual measured properties. *PASS Online predicts over 4000 kinds of biological activity, including pharmacological effects, mechanisms of action, toxic and adverse effects, interaction with metabolic enzymes and transporters, influence on gene expression, etc.* (PASS).

In practice, however, *property deficit* means that the main scheme in PE analyses is usually replaced by architectures with a single property but measured for a large number of objects or even where a property that is probed for a small number of actual FCS compounds are then predicted in the VCS population. In such procedures, both the FCS and VCS compounds are coded by the same molecular

descriptors that are more or less relatively easy for *in silico* calculation, which means that they could be available inexpensively. Accordingly, in the next two paragraphs we will discuss two types of PE data with:

- large number of property annotations (PE-LPE) or even
- property prediction (PE-PPE).

4.3 Molecular Big Data by a Large Number of Objects with a Single Property Annotation (PE-LPA)

The expansion of the PE into big data using actual property measurements for large FCS domains has been known for years. This includes screening. We can identify this approach in the screening of random compounds versus different biological targets. Some important drugs have been found by screening, e.g. the diketo acid (DKA) drugs against HIV integrase (Koehler 2000). More or less combinatorial chemistry also targets the property measurements of random compounds (Schneider 2002; Teague et al. 1999). Formally, we increase the number of property data simply by the increase of the number of objects for which a single property was usually measured. In other words, if the *height* of a single property column increases, the number of property measurements also increases. However, public access to the data from random screening or combinatorial chemistry of pharma R&D is limited.

A *property deficit* can be better understood if we realize that in recent years all of the biological activity data that has been published in the literature, e.g., Journal of Medicinal Chemistry has been carefully analyzed (Walters et al. 2011). This indicates the specific quality of this data. In particular, journals are only accepting reports that have positive results. This means that the mean values of biological activity of the drug candidate that are published in medicinal chemistry are skewed into higher numbers in comparison to the activity of actual drugs, while low activity data is less available. In one of the largest study that used more than three million structure-activity relationship data points that were collected for 898 human targets that compared marketed drugs and clinical candidates with bioactive compounds, some problems with optimizing pIC₅₀ were addressed (Tyrchan et al. 2009). Can pIC₅₀ data be used to model and analyze the fate of drug candidates and drugs on the market? An interesting study of the pIC₅₀ values for all drugs versus those at the top bestseller list was recently reported (Polanski et al. 2016b). However, does the inclusion of all of the available data mean that we have big data? The answer could be *yes* if this is determined by data complexity and *no* when the number of data entries is considered.

The classical quantitative structure activity/property relationship (QSAR/QSPR) still focuses on a relatively small series of compounds. At the same time, the development of combinatorial synthesis has not only led to large collections of chemical compounds but has also substantially enriched the market of chemical

reagents and building blocks. In fact, ready-to-use building block libraries of millions of compounds are now offered on the market. This provides an interesting source of big data annotated with *prices* that can be interpreted as the economic property of chemical compounds. Recently, quantitative structure-economic relationships (QSER) for a large dataset of a commercial building block library of over 2.2 million chemicals have been analyzed. It appears that on average what we are paying for is a quantity of matter. On the other hand, the influence of the synthetic availability scores is also revealed. Finally, we are buying substances by looking at molecular graphs or molecular formulas. Thus, those molecules that have a higher number of atoms look more attractive and are, on average, also more expensive (Polanski et al. 2016a).

Fialkowski et al. (2005) and Grzybowski et al. (2009) in the broadest structure-property studies ever reported investigated all Beilstein registered chemical compounds to analyze the *architecture and evolution of organic chemistry* (AEOC). The structure was here represented by the MW value, while the property—by chemical reactivity, i.e., the number of the entries of the given compound as the *reagent* or *product* in the reaction database, respectively. Probing the distribution of the MWs of the Beilstein registered compounds is another interesting example of the AEOC analysis. As we discussed above MW can be both the descriptor (if calculated) and the property (if measured). As we know, however, in the AEOC model all compounds were really synthesized, thus, represented here the FCS subspace. Therefore, their MWs *could be measured*, which means that in fact for some of these compounds the MW has been measured, e.g., by the MS spectroscopy. For the other, the MWs were obviously predicted from the other measurements, e.g., NMR, X-ray structures, etc. The credibility of such MW prediction is very high and the only exceptions are the errors in the structure determination. Therefore, formally the MW distribution analysis is somewhere in between PE-LPE analysis and the PE with predicted property annotation that is analyzed in Sect. 4.4.

Technically, data binning was used in both in the WUOC or the QSER studies. In Fig. 5 different versions of binning opportunities are analyzed while Fig. 6 illustrates the QSER and Fig. 7—AEOC model, respectively. Obviously, the analysis of million of observables must provide the representation that is much more superficial but at the same time more general than the small data analyses. Accordingly, this means higher complexity but lower precision. The term molecular statistics was coined to describe such a probe. *We use here molecular classes more often than individual compounds massive probing of the formal structure of chemical space, by sampling chemical compounds to measure the statistics of molecular descriptors and/or properties. Obviously, molecular statistics is a type of SAR, but the difference is that we focus here more on the compound classes that are more fuzzy than on individual drugs or chemotypes, e.g., all drugs, non-drugs, FDA approvals, etc.* (Polanski et al. 2016a).

An interesting statistical verification of the binning in molecular analyses can be found in Polanski et al. (2016a) and Kenny and Montanari (2013). While Kenny indicated the statistical malfunction of binning by the substantial signal

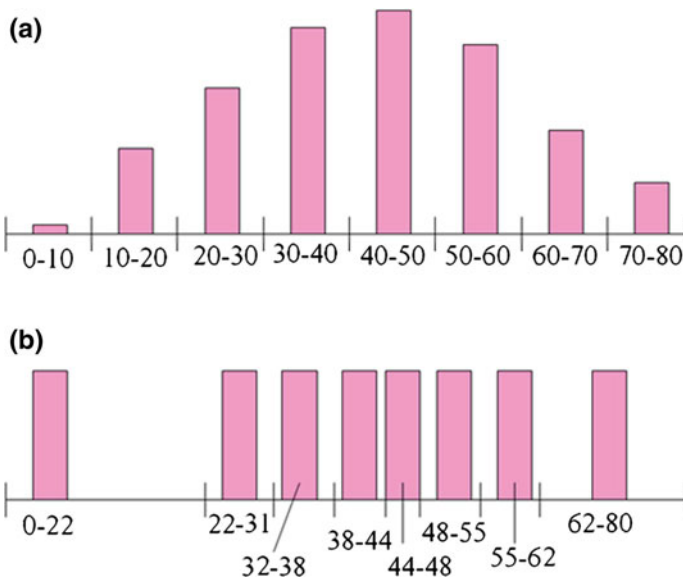


Fig. 5 Equi-width (a) and depth (b) binning schemes. Modified from (Kulkarni)

noise reduction, Polanski et al. have shown that the characteristics of the informative molecular statistics is significantly different from this of the random signals (Fig. 6b vs. c).

4.4 Molecular Big Data by the Increase of Objects with Predicted Property Annotation (PE-PPA)

The lack of activity data means we are frequently attempting to predict a property for a large chemical compound set on the basis of a relatively small compound series. Obviously, the complication here is perhaps even larger than in the records that are traditionally called big data. Therefore, we will call this type of information the statistics of the large number of objects with the expansion of the predicted property annotation (PPA).

The most obvious example is the prediction of the partition coefficient. Actually, partition coefficients have been measured experimentally for only 30,000 substances (Martel et al. 2013), a tiny fraction of the FCS population. Based on this measurement, we are estimating the logP for millions of virtual designed molecules or not measured substances, because the actual partition coefficient values will be measured for chemical compounds only in rare cases, even if they have been synthesized. Formally, this operation can be classified as the mapping of partition coefficients into logP (both for unmeasured FCS and unavailable VCS compounds).

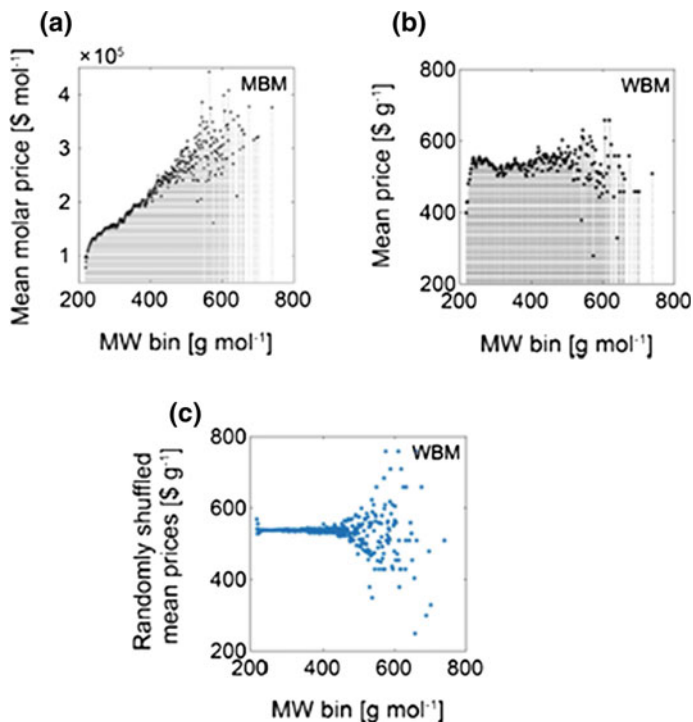


Fig. 6 The big data quantitative structure-price analysis of the building block library of ca. 2.5 million chemical compounds. The prices shown as a function of MW bins. On average we are paying here for the sample weight (a), the prices on the weight basis, WBM, b compared to the randomly shuffled prices c indicating substantial differences, which proves the information extracted from the model is not a random noise. Modified from Polanski et al. (2016a)

For a number of substances this means a prediction of the partition coefficient property based on the calculation of the log P value. The *additivity concept* allows to calculate the logP by summing the contributions from structure fragments. Technically, individual contributions are calculated from the regression model that relates the partition coefficient measured for the substances to fragmental molecular descriptors in the FCS. The logP value for the whole molecule can then be calculated by simply putting together the increments that describe the individual molecular fragments. This operation that can be performed for each molecule in the whole chemical space means we are predicting partition coefficient based on the log P versus partition coefficient regression. In particular, a number of different calculation systems have been developed here, e.g., the fragmental Rekker algorithm (Rekker 1977). Statistically, prediction is prone to failures; however, from the chemical point we can explain the problems with prediction accuracy if we understand that we are replacing here a substance (in vitro) with an isolated molecule (in silico) and then come back to the substance (in vitro) after its actual synthesis.

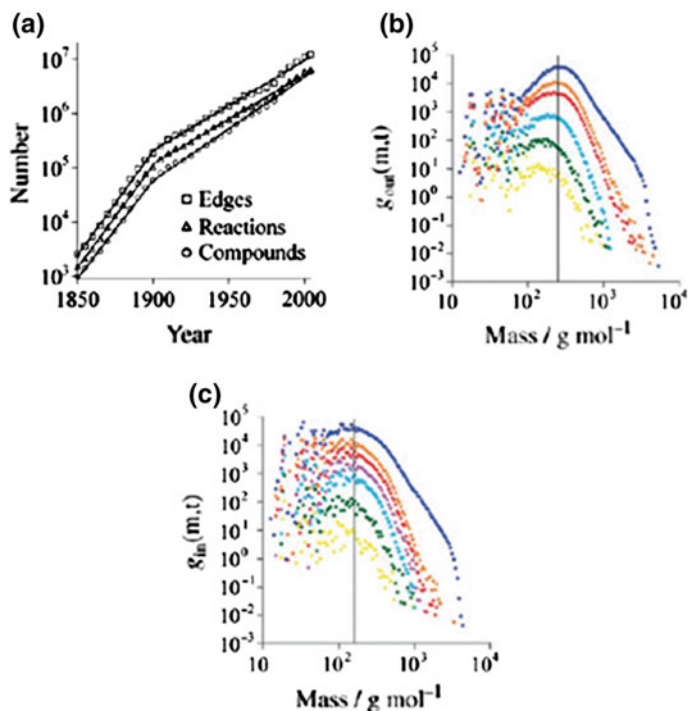


Fig. 7 Representative AEOC models. The number of edges, reactions and compounds for the network linking reagents with products versus publication year (a) and the distribution of the MW (g/mol) for the products (b) and reagents (c), respectively. The data are binned in years (a) or MWs (b, c), respectively. Modified from Fialkowski et al. (2005)

Another example of the PE-PPA model is the application of molecular data in virtual screening in which a small series of compounds of a known structure and activity can be used to screen a million populations of VCS compounds that are represented only by descriptors in order to predict their potential activity. Let us take an example where 1.5 million databases of the available commercial compounds were screened in the search for potential new HIV1 active chemotypes (Kurczyk et al. 2015). Thus, a set of 1,140 compounds with a determined HIV-1 IN inhibition was fetched from the ChEMBL v.12 database (ChEMBL). The library of these compounds was further divided according to their HIV-1 IN inhibition ranges (Fig. 8) to obtain the subsets of actives and inactives. The Klekota-Roth fingerprint (Klekota and Roth 2008) was used as the molecular descriptor representation for developing the classification models. The HIV-specific privileged fragment descriptors were identified in the ChEMBL-EBI compound subset with a known activity and used for screening using the weight-based scoring function method. Actually, the new chemotype hits when the tested compound appeared to indicate a low anti-HIV activity.

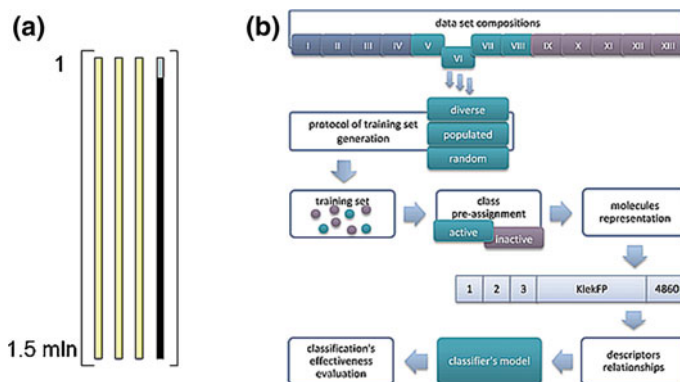


Fig. 8 The big data PE-PPA matrix with the majority of property annotation data predicted (*black*) on the basis on the relatively small number of real measurements (*blue*) (a) and complex analysis scheme (b). Details in text. Modified from Kurczyk et al. (2015)

5 Conclusions

We discussed what big data is and how important big data can be in drug design. In particular, we analyzed the big data types that are available, in particular, in structure-property studies. We also indicated several basic big molecular data types. Basically, these are the big data that are generated by descriptor (DE) or property (PE) expansion. Actually, however, measured property data are seldom available, in other words, we are under a *property deficit*. *Property deficit* means that the PE data are usually getting bigger not by the number of property types but by an increase in the number of chemical compounds that are annotated with a single property type. This also means that property prediction data is an important architecture in the big structure-property data. At the same time, the *property deficit* is among the main obstacles that limit the application of big data in structure-property studies.

References

- Blum, L. C., & Reymond, J.-L. (2009). 970 million drug like small molecules for virtual screening in the chemical universe database GDB-13. *Journal of the American Chemical Society*, *131*, 8732–8733.
- CAS, 100-millionth-substance. <https://www.cas.org/news/media-releases/100-millionth-substance>.
- CDD, Collaborative Drug Discovery. www.collaborativedrug.com.
- ChEMBL, ChEMBL. <http://www.ebi.ac.uk/chembl>.
- Cheng, T., Pan, Y., Hao, M., Wang, Y., & Bryant, S. H. (2014). PubChem applications in drug discovery: A bibliometric analysis. *Drug Discovery Today*, *19*, 1751–1756.
- Consonni, V., & Todeschini, R. (2010). Molecular descriptors. In T. Puzyn et al. (Eds.), *Recent advances in QSAR studies* (pp. 29–102). Springer Science + Business Media B.V.

- Cramer, R. D., Patterson, D. E., & Bunce, J. D. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society*, *110*, 5959–5967.
- Cukier, K., & Mayer-Schönberger, V. (2013). *Big Data: A revolution that will transform how we live, work, and think*. New York: Business & Economics.
- Fialkowski, M., Bishop, K. J. M., Chubukov, V. A., Campbell, C. J., & Grzybowski, B. A. (2005). Architecture and evolution of organic chemistry. *Angewandte Chemie (International ed. in English)*, *44*, 7263–7269.
- Gasteiger, J. (Ed.). (2003). *Handbook of chemoinformatics: From data to knowledge, 4 volumes*. Weinheim: Wiley-VCH.
- Gasteiger, J., & Engel, T. (Eds.). (2003). *Chemoinformatics: A textbook*. Weinheim: Wiley-VCH.
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*, 1012–1014.
- Grzybowski, B. A., Bishop, K. J. M., Kowalczyk, B., & Wilmer, C. E. (2009). The 'wired' universe of organic chemistry. *Nature Chemistry*, *1*, 31–36.
- IUPAC, Goldbook. <http://goldbook.iupac.org>.
- Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.
- Kenny, P. W., & Montanari, C. A. (2013). Inflation of correlation in the pursuit of drug-likeness. *Journal of Computer-Aided Molecular Design*, *27*, 1–13.
- Klekota, J., & Roth, F. P. (2008). Chemical substructures that enrich for biological activity. *Bioinformatics*, *24*, 2518–2525.
- Koehler, Ch. S. W. (2000). AIDS, arteries and engineering, epidemics end entrepreneurs: In the pharmaceutical century: Ten decades of drug discovery, ACS. <http://www3.uah.es/farmamol/The%20Pharmaceutical%20Century/Ch7.html>.
- Kulkarni, S. Introduction to data mining. <http://www.slideshare.net/sushil.kulkarni/ch-1-intro-to-data-mining-presentation>.
- Kurczyk, A., Warszycki, D., Musiol, R., Kafel, R., Bojarski, A. J., & Polanski, J. (2015). Ligand-based virtual screening in a search for novel anti-HIV-1 chemotypes. *Journal of Chemical Information and Modeling*, *55*, 2168–2177.
- Laney, D. (2016). 3-D data management: Controlling data volume, velocity. <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>.
- Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: Traps in Big Data analysis. *Science*, *343*, 1203–1205.
- Ldtopology, Low Dimensional Topology. ldtopology.wordpress.com.
- Lipinski, C. A. (2000). Drug-like properties and the causes of poor solubility and poor permeability. *Journal of Pharmacological and Toxicological Methods*, *44*, 235–249.
- Lipinski, C., & Hopkins, A. (2004). Navigating chemical space for biology and medicine. *Nature*, *432*, 855–861.
- Maheshwari, A. (2014). *Data analytics made accessible* (Kindle edition Amazon).
- Martel, S., Gillerat, F., Carosati, E., Maiarelli, D., Tetko, I. V., Mannhold, R., et al. (2013). Large, chemically diverse dataset of logP measurements for benchmarking studies. *European Journal of Pharmaceutical Sciences*, *48*, 21–29.
- Musiol, R., Jampilek, J., Kralova, K., Richardson, D. R., Kalinowski, D., Podeszwa, B., et al. (2007). Investigating biological activity spectrum for novel quinoline analogues. *Bioorganic & Medicinal Chemistry*, *15*, 1280–1288.
- PASS, Prediction of Activity Spectra for Substances. www.pharmaexpert.ru/passonline/.
- Polanski, J. (2009a). *Chemoinformatics*. In S. D. Brown, R. Tauler, & B. Walczak (Eds.), *Comprehensive chemometrics*. Elsevier.
- Polanski, J. (2009b). Receptor dependent multidimensional QSAR for modeling drug—Receptor interactions. *Current Medicinal Chemistry*, *16*, 3243–3257.
- Polanski, J., & Gasteiger, J. (2016). Computer representation of chemical compounds. In J. Leszczynski (Ed.), *Handbook of computational chemistry*. Springer.

- Polanski, J., & Walczak, B. (2000). The comparative molecular surface analysis (COMSA): A novel tool for molecular design. *Computers & Chemistry*, *24*, 615–625.
- Polanski, J., Bogocz, J., & Tkocz, A. (2016a). The analysis of the market success of FDA approvals by probing top 100 bestselling drugs. *Journal of Computer-Aided Molecular Design*, *30*, 381–389.
- Polanski, J., Kucia, U., Duszkiewicz, R., Kurczyk, A., Magdziarz, T., & Gasteiger, J. (2016b). Molecular descriptor data explain market prices of a large commercial chemical compound library. *Scientific Reports*, *6*, 28521.
- Rekker, R. F. (1977). *The hydrophobic fragment constant*. New York, NY: Elsevier.
- Reymond, J. L. (2015). The chemical space project. *Accounts of Chemical Research*, *48*, 722–730.
- Rosenblum, B., & Kuttner, F. (2011). *Quantum enigma: Physics encounters consciousness*. Oxford University Press.
- Schneider, G. (2002). Trends in virtual combinatorial library design. *Current Medicinal Chemistry*, *9*, 2095–2101.
- Szlezák, N., Evers, M., Wang, J., & Pérez, L. (2014). The role of big data and advanced analytics in drug discovery, development, and commercialization. *Clinical Pharmacology and Therapeutics*, *95*, 492–495.
- Teague, S. J., Davis, A. M., Leeson, P. D., & Oprea, T. (1999). The design of lead like combinatorial libraries. *Angewandte Chemie International Edition*, 3743–3748.
- Todeschini, R., & Consonni, V. (2000). *Handbook of molecular descriptors*. Weinheim: Wiley-VCH.
- Tyrchan, C., Blomberg, N., Engkvist, O., Kogej, T., & Muresan, S. (2009). Physicochemical property profiles of marketed drugs, clinical candidates and bioactive compounds. *Bioorganic & Medicinal Chemistry Letters*, *19*, 6943–6947.
- Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., & Taylor, R. D. (2003). Improved protein-ligand docking using GOLD. *Proteins*, 609–623.
- Walters, W. P., Green, J., Weiss, J. R., & Murcko, M. A. (2011). What do medicinal chemists actually make? A 50-year retrospective. *Journal of Medicinal Chemistry*, *54*, 6405–6416.
- Wang, Y., Suzek, T., Zhang, J., Wang, J., He, S., Cheng, T., et al. (2014). *Nucleic Acids Research*, *42*, D1075-82.

Index

A

Activity cliffs, 303, 309–322, 326–331
Activity landscape, 309, 310, 312, 313, 315–320, 327–329
Acute oral toxicity, 127, 128
ADME properties predictions, 354, 357, 362, 370
Affinity predictions, 368
Agriculture, 204, 207, 211, 222, 227, 228, 232, 280, 293
Agrochemical, 203, 205, 211, 222, 225, 226, 278, 280, 293
ANOVA, 89, 96, 98, 101, 103
Antagonists of fibrinogen receptor, 122
Antioxidants, 203, 204, 215–220, 269–271, 273–275

B

Big data, 529–533, 536, 539–541, 543, 546, 547, 550
Bioaccumulation refinement, 426, 427, 438
Bioactive peptides, 204, 205
Biodescriptors, 77
Biotransformation prediction, 428, 430, 431, 436, 438, 440
Black box learning methods, 14, 28
Bromodomains, 303, 304, 306, 323, 328, 331
Business intelligence, 531

C

Carcinogenicity, 503–510, 512, 513, 518–523
Chemical databases, 536, 540
Chemical similarity, 152
Chemical space, 530, 533, 535, 542, 546, 548
Chemical space mapping, 172
Cheminformatics, 7
Chemogenomics, 6
Chromone, 272
Clique, 151, 152, 154

CoMFA, 228, 232, 238, 242, 244, 246, 250, 251, 254, 257, 260, 265, 267, 274, 539, 543, 544
CoMSIA, 228, 232, 242, 246, 254, 257, 267, 272, 273, 276, 277
CORAL, 453, 455–457, 459, 462, 463, 467, 468
Coumarin, 222, 273
Counter propagation artificial network (CP ANN), 503, 506, 507, 511–513, 516, 518, 521
Critical properties, 120, 121
Cross-validation, 89–91, 93, 95, 97–99, 103

D

Data analysis, 540
Database, 203, 226, 234, 249, 251, 255, 266, 277–280
Data binning, 541, 542, 546
Data management, 540
Data mining, 34, 391
Data populations in chemistry, 533
Data preprocessing, 89, 93, 97, 98, 100, 101
DNA methyltransferase, 303, 304, 327
Docking, 203, 228, 239, 240, 242, 249, 256, 260, 287
Drug design, 4–7, 25, 27
Drug discovery, 4, 7, 15, 17, 20, 45, 46
Drug-likeness, 347, 352, 372

E

Economics, 530, 531, 535
Electrotopological state, 57, 60, 63, 75–77
Epigenetics, 304, 305, 307, 323, 326
Epi-informatics, 303
European Union, 223, 229, 230
Expert system, 234, 286, 287
External validation, 89–93, 97, 99, 101–103

F

- Fish toxicity, 459, 460, 464, 465
- Flavonoid, 217, 219, 269, 271
- Flavor, 203, 204, 221, 222, 264, 265, 278, 279
- Food, 203–205, 215, 217, 220–226, 264, 266, 278–280, 286
- Food supplements, 203–205, 215, 216, 220, 260, 264
- Fragrance, 204, 268, 279
- Free radical, 203, 205, 216–218, 220, 269, 270, 272, 275, 277
- Free-Wilson models, 118
- Fungicides, 211, 213, 245

G

- Generative topographic maps, 169, 196
- GFA, 259, 267
- Global models, 341, 345, 357, 367, 371
- Google algorithm for spread of flu prediction, 531
- Graph theory, 154
- Growth regulators, 204, 215, 216, 244, 258

H

- Hammett constants, 117
- HDAC, 304, 306, 323, 324, 326, 331
- Herbicide, 211, 232, 237, 239, 241–245, 280
- Hierarchy of scientific explanation, 530
- Hostility to topological indices, 79

I

- Indicators, 91, 96, 97, 99
- Information content, 57, 60, 63, 65
- Insecticide, 211, 214, 249, 252, 255, 257
- In silico, 203, 204, 215, 216, 225–230, 252, 261, 268, 270, 278, 287, 293, 425, 427–429, 442, 444, 445
- In silico models, 503, 504, 522
- Inverse QSAR, 75
- IVIVE, 426–428, 430, 442

K

- Kohonen Self-Organizing Maps, 168, 169

L

- Local models, 339, 341, 347, 357, 361, 371
- LogP, 534, 547, 548

M

- Machine learning, 4, 5, 7, 8, 14, 15, 18, 28, 30, 33, 40, 46
- Machine learning for nanomaterial data, 390, 397

- Maximum Common Substructure (MCS), 150, 152, 154–163
- Merits for QSAR, 91
- Metabolism, 427, 442
- MLR, 238, 239, 244, 246, 249, 252, 265, 267–269, 274
- Model interpretability, 341, 344
- Models for regulatory use, 522
- Molecular big data architectures, 542
- Molecular big data by a large number of objects with a single property annotation (PE-LPA), 545
- Molecular big data by descriptor expansion (DE), 543
- Molecular big data by property expansion (PE), 544
- Molecular big data by the increase of objects with predicted property annotation (PE-PPA), 547
- Molecular connectivity, 61, 69, 70, 79
- Molecular descriptor, 4, 6–8, 14, 16–20, 22, 27, 46, 530, 534–536, 539, 540, 542, 545, 546, 548, 549
- Multi-objective drug discovery, 342

N

- Nanomaterial characterization, 386
- Non-congeneric chemicals, 503, 506

O

- OECD, 224, 229–231, 233, 234, 252, 287, 293

P

- PASS, 544
- Pesticide, 211, 223, 224, 248, 249, 251, 252, 280, 287, 471–492
- Pharmacophore, 203, 205, 227, 252–254, 260, 272, 273, 275, 276
- Physico-chemical/structural properties, 386–390, 392, 393, 395, 396, 399–401, 404, 408, 413, 417
- Phytochemical, 204, 215–217, 222, 225, 226, 278, 286, 293
- PLS, 239, 249, 250, 266, 271, 272, 275
- Prediction, 472–474, 484–491
- Property, 529, 534–536, 540, 542–548, 550
- Property deficit, 529, 544, 545, 550
- Proteochemometric model, 6

Q

- QSARINS, 433–435, 437
- QSAR model interpretation, 131
- QSAR models, 472, 474, 484–492
- Quail toxicity, 458, 465, 467

- Quantitative Structure-Activity Relationship (QSAR), 4–8, 10, 14–20, 22, 23, 26, 31, 33, 38, 40, 45, 46, 189, 205, 224, 226–229, 232, 234, 238, 239, 241, 244, 245, 247–275, 277, 278, 286, 287, 293, 385, 412, 453, 455–457, 459, 462–464, 466–468, 503, 505, 509, 516, 518, 522, 523, 530, 533, 534, 536, 539, 543
- Quantitative structure-economics relationship, 531
- Quantitative Structure-Property Relationship (QSPR), 4, 10, 17, 530, 534
- R**
- Randić, 57, 61, 62, 68, 69, 71, 72, 74, 77, 79
- Rat toxicity, 454, 456, 458, 465–467
- REACH, 226, 229, 287, 503–505, 507, 509, 522
- Read-across, 149, 150, 152–154, 163
- Responsibility patterns, 172–177, 188–196, 197
- S**
- SEARS, 306, 307, 331
- Simplex Representation of Molecular Structure (SiRMS), 107–111, 113, 115, 134, 137, 138, 142, 144
- Software, 59–61, 78
- Stereochemical interpretation of QSAR models, 109
- Structural alerts, 153, 154, 163
- Structural and physicochemical interpretation of QSAR models, 109, 112, 142
- Sum of ranking differences, 91, 93, 95–103
- T**
- Toxicity, 471, 473, 479, 484–489, 491, 492
- U**
- US EPA, 223, 224, 249, 250, 287
- US FDA, 221, 223, 225, 226, 279
- V**
- Virtual screening, 203, 227, 242, 287
- Virucide, 215, 258
- W**
- White box learning methods, 37
- Wiener, 63, 64