

Challenges and Advances
in Computational Chemistry and Physics 33
Series Editor: Jerzy Leszczynski

Alla P. Toropova
Andrey A. Toropov *Editors*

QSPR/QSAR Analysis Using SMILES and Quasi-SMILES

MOREMEDIA



Springer

Challenges and Advances in Computational Chemistry and Physics

Volume 33

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.


All contributions to edited volumes should undergo standard peer review to ensure high scientific quality, while monographs should be reviewed by at least two experts in the field. Submitted manuscripts will be reviewed and decided by the series editor, Prof. Jerzy Leszczynski.


Alla P. Toropova · Andrey A. Toropov
Editors

QSPR/QSAR Analysis Using SMILES and Quasi-SMILES

 Springer

Editors

Alla P. Toropova 
Department of Environmental Health
Science
Institute of Pharmacological Research
Mario Negri IRCCS
Milan, Italy

Andrey A. Toropov 
Department of Environmental Health
Science
Institute of Pharmacological Research
Mario Negri IRCCS
Milan, Italy

ISSN 2542-4491

ISSN 2542-4483 (electronic)

Challenges and Advances in Computational Chemistry and Physics

ISBN 978-3-031-28400-7

ISBN 978-3-031-28401-4 (eBook)

<https://doi.org/10.1007/978-3-031-28401-4>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are solely and exclusively licensed 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, expressed 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.

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

Preface

Who is this book for intended? Primarily for students who are planning their carrier. Ph.D. students can also get valuable ideas for their careers if they are sure that their scientific activity somehow connects with chemistry, biology, medicine, informatics, and mathematical chemistry. The author's team contains specialists in different directions of chemistry, biochemistry, and medicinal chemistry. The geography of the authors is vast enough: USA, Canada, Iran, India, China, Uzbekistan, Czech Republic, Portugal and Italy.

It seems that recognizing the differences in the paths of transition of randomness into regularity or, conversely, the ways of randomness into stable chaos may be of interest to everyone since this task affects any area of human activity. In fact, this book describes attempts to solve the mentioned problem concerning development processes QSPR/QSAR and nano-QSPR/QSAR.

The curious intrigue of the proposed book demonstrates the ability of randomness to provide patterns through variational autoencoders (VAEs) defined over SMILES string and molecular graph, the Monte Carlo technique, and using so-called quasi-SMILES (i.e., traditional SMILES extended via special symbols which are reflecting experimental conditions). However, the philosophic principle "nothing is the only" should make the reader sure that every model should be validated as much as possible, i.e., checked up under a diversity of experimental conditions.

Thus, there is the probability that the book can become curiously and attractive to various "random" readers (professors, engineers, players) who are capable of curios and wonder relevant to the process of building up models for different phenomena.

Milan, Italy

Alla P. Toropova
Andrey A. Toropov

Contents

Part I Theoretical Conceptions

- 1 Fundamentals of Mathematical Modeling of Chemicals Through QSPR/QSAR** 3
Andrey A. Toropov, Maria Raskova, Ivan Raska Jr.,
and Alla P. Toropova
- 2 Molecular Descriptors in QSPR/QSAR Modeling** 25
Shahin Ahmadi, Sepideh Ketabi, and Marjan Jebeli Javan
- 3 Application of SMILES to Cheminformatics and Generation of Optimum SMILES Descriptors Using CORAL Software** 57
Andrey A. Toropov and Alla P. Toropova

Part II SMILES Based Descriptors

- 4 All SMILES Variational Autoencoder for Molecular Property Prediction and Optimization** 85
Zaccary Alperstein, Artem Cherkasov, and Jason Tyler Rolfe
- 5 SMILES-Based Bioactivity Descriptors to Model the Anti-dengue Virus Activity: A Case Study** 117
Soumya Mitra, Sumit Nandi, Amit Kumar Halder,
and M. Natalia D. S. Cordeiro

Part III SMILES for QSPR/QSAR with Optimal Descriptors

- 6 QSPR Models for Prediction of Redox Potentials Using Optimal Descriptors** 139
Karel Nesměrak and Andrey A. Toropov
- 7 Building Up QSPR for Polymers Endpoints by Using SMILES-Based Optimal Descriptors** 167
Valentin O. Kudyshkin and Alla P. Toropova

Part IV Quasi-SMILES for QSPR/QSAR

- 8 Quasi-SMILES-Based QSPR/QSAR Modeling** 191
Shahin Ahmadi and Neda Azimi
- 9 Quasi-SMILES-Based Mathematical Model for the Prediction of Percolation Threshold for Conductive Polymer Composites** 211
Swayam Aryam Behera, Alla P. Toropova, Andrey A. Toropov, and P. Ganga Raju Achary
- 10 On the Possibility to Build up the QSAR Model of Different Kinds of Inhibitory Activity for a Large List of Human Intestinal Transporter Using Quasi-SMILES** 241
P. Ganga Raju Achary, P. Kali Krishna, Alla P. Toropova, and Andrey A. Toropov
- 11 Quasi-SMILES as a Tool for Peptide QSAR Modelling** 269
Md. Moinul, Samima Khatun, Sk. Abdul Amin, Tarun Jha, and Shovanlal Gayen

Part V SMILES and Quasi-SMILES for QSPR/QSAR

- 12 SMILES and Quasi-SMILES Descriptors in QSAR/QSPR Modeling of Diverse Materials Properties in Safety and Environment Application** 297
Yong Pan, Xin Zhang, and Juncheng Jiang
- 13 SMILES and Quasi-SMILES in QSAR Modeling for Prediction of Physicochemical and Biochemical Properties** 327
Siyun Yang, Supratik Kar, and Jerzy Leszczynski

Part VI Possible Ways of Nano-QSPR/Nano-QSAR Evolution

- 14 The CORAL Software as a Tool to Develop Models for Nanomaterials' Endpoints** 351
Alla P. Toropova and Andrey A. Toropov
- 15 Employing Quasi-SMILES Notation in Development of Nano-QSPR Models for Nanofluids** 373
Kimia Jafari and Mohammad Hossein Fatemi

Part VII Possible Ways of QSPR/QSAR Evolution in the Future

16 On Complementary Approaches of Assessing the Predictive Potential of QSPR/QSAR Models	397
Andrey A. Toropov, Alla P. Toropova, Danuta Leszczynska, and Jerzy Leszczynski	
17 CORAL: Predictions of Quality of Rice Based on Retention Index Using a Combination of Correlation Intensity Index and Consensus Modelling	421
Parvin Kumar and Ashwani Kumar	
Index	463

Contributors

Sk. Abdul Amin Natural Science Laboratory, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, West Bengal, India;
Department of Pharmaceutical Technology, JIS University, Agarpara, Kolkata, West Bengal, India

P. Ganga Raju Achary Department of Chemistry, Institute of Technical Education and Research (ITER), Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India

Shahin Ahmadi Department of Chemistry, Faculty of Pharmaceutical Chemistry, Tehran Medical Sciences, Islamic Azad University, Tehran, Iran

Zaccary Alperstein Variational AI, Vancouver, BC, Canada

Neda Azimi Advanced Chemical Engineering Research Center, Razi University, Kermanshah, Iran

Swayam Aryam Behera Department of Chemistry, Institute of Technical Education and Research (ITER), Siksha 'O' Anusandhan University, Bhubaneswar, Odisha, India

Artem Cherkasov Vancouver Prostate Centre, UBC, Vancouver, BC, Canada

M. Natalia D. S. Cordeiro LAQV@REQUIMTE, Faculty of Sciences, University of Porto, Porto, Portugal

Mohammad Hossein Fatemi Chemometrics Laboratory, Faculty of Chemistry, University of Mazandaran, Babolsar, Iran

Shovanlal Gayen Laboratory of Drug Design and Discovery, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, West Bengal, India

Amit Kumar Halder Dr. B. C. Roy College of Pharmacy and Allied Health Sciences, Durgapur, West Bengal, India;
LAQV@REQUIMTE, Faculty of Sciences, University of Porto, Porto, Portugal

Kimia Jafari Chemometrics Laboratory, Faculty of Chemistry, University of Mazandaran, Babolsar, Iran

Marjan Jebeli Javan Department of Chemistry, Faculty of Pharmaceutical Chemistry, Tehran Medical Sciences, Islamic Azad University, Tehran, Iran

Tarun Jha Natural Science Laboratory, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, West Bengal, India

Juncheng Jiang College of Safety Science and Engineering, Nanjing Tech University, Nanjing, China

Supratik Kar Chemometrics and Molecular Modeling Laboratory, Department of Chemistry, Kean University, Union, NJ, USA

Sepideh Ketabi Department of Chemistry, Faculty of Pharmaceutical Chemistry, Tehran Medical Sciences, Islamic Azad University, Tehran, Iran

Samima Khatun Laboratory of Drug Design and Discovery, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, West Bengal, India

P. Kali Krishna Department of Bioinformatics, B.J.B Autonomous College, Bhubaneswar, Odisha, India

Valentin O. Kudyshkin Institute of Polymer Chemistry and Physics, Academy of Sciences of the Republic of Uzbekistan, Tashkent, Uzbekistan

Ashwani Kumar Department of Pharmaceutical Sciences, Guru Jambheshwar University of Science and Technology, Hisar, Haryana, India

Parvin Kumar Department of Chemistry, Kurukshetra University, Kurukshetra, Haryana, India

Danuta Leszczynska Department of Civil and Environmental Engineering, Interdisciplinary Nanotoxicity Center, Jackson State University, Jackson, MS, USA

Jerzy Leszczynski Department of Chemistry, Physics and Atmospheric Sciences, Interdisciplinary Center for Nanotoxicity, Jackson State University, Jackson, MS, USA

Soumya Mitra Dr. B. C. Roy College of Pharmacy and Allied Health Sciences, Durgapur, West Bengal, India

Md. Moinul Laboratory of Drug Design and Discovery, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, West Bengal, India

Sumit Nandi Dr. B. C. Roy College of Pharmacy and Allied Health Sciences, Durgapur, West Bengal, India

Karel Nesměřák Department of Analytical Chemistry, Faculty of Science, Charles University, Prague 2, Czech Republic

Yong Pan College of Safety Science and Engineering, Nanjing Tech University, Nanjing, China

Ivan Raska Jr. 3rd Medical Department, 1st Faculty of Medicine, Charles University in Prague, Prague 2, Czech Republic

Maria Raskova 3rd Medical Department, 1st Faculty of Medicine, Charles University in Prague, Prague 2, Czech Republic

Jason Tyler Rolfe Variational AI, Vancouver, BC, Canada

Andrey A. Toropov Laboratory of Environmental Chemistry and Toxicology, Department of Environmental Health Science, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milano, Italy

Alla P. Toropova Laboratory of Environmental Chemistry and Toxicology, Department of Environmental Health Science, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milano, Italy

Siyun Yang Chemometrics and Molecular Modeling Laboratory, Department of Chemistry, Kean University, Union, NJ, USA

Xin Zhang College of Safety Science and Engineering, Nanjing Tech University, Nanjing, China

Abbreviations

AAD	Average absolute deviation
ACE	Angiotensin-converting enzymes
AD	Applicability domain
AFM	Atomic force microscopy
ANFIS	Adaptive neuro-fuzzy inference system
ANN	Artificial neural networks
AZI	Augmented Zagreb index
BET	Brunauer, Emmett and Teller
CCC	Concordance correlation coefficient
CII	Correlation intensity index
CORAL	Correlation and logic
C_p	Isobaric heat capacity
CW	Correlation weights
DCW	Descriptor of correlation weights
DHFR	Dihydrofolate reductase
DLS	Dynamic light scattering
DTR	Decision tree regression
EDX	Energy dispersive X-ray spectrometry
EG	Ethylene glycol
EM	Electronic microscopy
EP	Endpoint
ESEM	Environmental scanning electron microscopy
F	Fischer ratio
FF-ANNs	Feed-forward artificial neural networks
FFF	Field flow filtration
FMO	Frontier molecular orbital theory
GBR	Gradient boosting regression
GNPs	Gold nanoparticles
GRNNs	Generalized regression neural networks
GRUs	Gated recurrent units
HOMO	Highest occupied molecular orbital

HSG	Hydrogen-suppressed molecular graphs
ICP-MS	Inductively coupled plasma mass spectrometry
ICPOES	Inductively coupled plasma emission spectroscopy
IIC	Index ideality of correlation
ILs	Ionic liquids
LC	Liquid chromatography
LDM	Liquid drop model
logP	Decimal logarithm of octanol-water partition coefficient
LSSVM	Least square support vector machine
LSTM	Long short-term memory
LUMO	Lowest unoccupied molecular orbital
MAE	Mean absolute error
MLP	Multilayer perceptron
MLR	Multiple regression analysis
MO-NPs	Metal oxide nanoparticles
MoRSE	3D-Molecular representation of structures based on electron diffraction
MVC	Multivariate characterization
MW	Molecular weight
MWCNTs	Multiwalls carbon nanotubes
NPs	Nanoparticles
OECD	Organization of Economic Co-operation and Development
PCA	Principal component analyses
PLS	Partial least-squares regression analysis
PPs	Principal properties
Q^2	Leave-one-out cross-validated correlation coefficient
QED	Quantitative estimate of drug-likeness
QSAR	Quantitative structure–activity relationship
QSGFEAR	Gibb’s free energy of activation relationship
QSPR	Quantitative structure–property relationship
Quasi-SMILES	Quasi-simplified molecular input-line entry-system
R^2	Determination coefficient (or squared correlation coefficient)
RBF	Radial basis function
RF	Random forest
RMSE	Root-mean-square error
RNNs	Recurrent neural networks
SA	SMILES attributes
SADT	Self-accelerating decomposition temperature
SFS	Sequential forward selection
SMILES	Simplified molecular input-line entry-system
SNN	Siamese neural network
SVM	Support vector machine
SVR	Support vector regression
SWCNTs	Single-wall carbon nanotubes
TC	Thermal conductivity

TEM	Transmission electron microscopy
TF	Target function
TMACC	Topological maximum cross-correlation
VAEs	Variational autoencoders
VIF	Variation inflation factor
WW	Hyper-Wiener index
ΔG^\ddagger	Gibb's activation free energy

Greek Symbols

ρ	Density
φ	Volume fraction of nanoparticle (%)

Subscripts

bf	Base fluid
nf	Nanofluid
p	Nanoparticle
v	Volume fraction

Chemical Formulas

Ag	Silver
Al_2O_3	Aluminum oxide
AlN	Aluminum nitride
Au	Gold
Bi_2O_3	Bismuth (III) oxide
CeO_2	Cerium (IV) oxide
Co_3O_4	Cobalt (II,III) oxide
Cr_2O_3	Chromium (III) oxide
Cu	Copper
CuO	Copper oxide
Dy_2O_3	Dysprosium (III) oxide
Fe	Iron
Fe_2O_3	Iron (III) oxide
Fe_3O_4	Iron (II,III) oxide
Gd_2O_3	Gadolinium (III) oxide
HfO_2	Hafnium (IV) oxide

In_2O_3	Indium (III) oxide
La_2O_3	Lanthanum oxide
MgO	Magnesium oxide
Mn_2O_3	Manganese (III) oxide
Mn_3O_4	Manganese (II,III) oxide
Ni_2O_3	Nickel (III) oxide
NiO	Nickel (II) oxide
Sb_2O_3	Antimony oxide
Si_3N_4	Silicon nitride
SiC	Silicon carbide
SiO_2	Silicon dioxide
SnO_2	Tin (IV) oxide
TiN	Titanium nitride
TiO_2	Titanium dioxide
WO_3	Tungsten (VI) oxide
Y_2O_3	Yttrium (III) oxide
Yb_2O_3	Ytterbium (III) oxide
ZnO	Zinc oxide
ZrO_2	Zirconium oxide

Part I
Theoretical Conceptions

Chapter 1

Fundamentals of Mathematical Modeling of Chemicals Through QSPR/QSAR



Andrey A. Toropov, Maria Raskova, Ivan Raska Jr., and Alla P. Toropova

Abstract The evolution of mathematical chemistry in its applications to establish the quantitative structure–property/activity relationships (QSPRs/QSARs) between molecular structure and the physicochemical and biochemical behavior of substances is discussed. The gradual improvement of molecular descriptors and the statistically validated methods developed for the above general task are described. The possible ways of applying and extending OECD principles are demonstrated via computational experiments to build QSPR/QSAR models. The leading role of validation in obtaining applicable models is noted. Stochastic procedures able to improve the reliability of QSPR/QSAR models are demonstrated.

Keywords Mathematical modeling · QSPR/QSAR · OECD principles · Molecular descriptors · Data curation · Reproducibility · Applicability domain · Model validation

1.1 Introduction

A considerable amount of valuable fundamental work on mathematical chemistry was carried out in the twentieth century and the first decade of the twenty-first century. However, this chapter will discuss the results obtained later, that is, in fact, in the second decade of the twenty-first century. Mathematical chemistry aims for many tasks. However, if try defines primary aims, one can extract the main word “model”. The term model itself relates to a grand manifold of phenomena. The impossibility

A. A. Toropov (✉) · A. P. Toropova
Laboratory of Environmental Chemistry and Toxicology, Department of Environmental Health Science, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156 Milano, Italy
e-mail: andrey.toropov@marionegri.it

M. Raskova · I. Raska Jr.
3rd Medical Department, 1st Faculty of Medicine, Charles University in Prague, U Nemocnice 1, 12808 Prague 2, Czech Republic

of connecting thinking experiments with traditional experiments becomes a great challenge that is the beginning of mathematical chemistry and other sciences. Science is the step from usual to unexpected. Naturally, mathematical chemistry is not an exception.

Thus, mathematical chemistry [1] is the area of research engaged in novel applications of mathematics to chemistry, biochemistry, and biology. A significant part of the above research is dedicated to the mathematical modeling of complex molecular phenomena that are “quite visible” at the macro-level and can be measured. Those are named endpoints (e.g., boiling point, heat capacity, or toxicity) [2]. Much has been said about the role and significance of the sciences, but if we single out the average, it turns out that few people are interested in this issue. Repeated sentences in the literature are necessary to find new words and meanings for old concepts. For instance, it has often been noted that science across all disciplines has become data-driven, leading to additional needs concerning software for collecting, processing, and analyzing data. Consequently, software becomes necessary for reproducibility and analysis of the evolution of scientific methods, often even in real time. Currently, research work is impossible without a computer for collecting, processing, and analyzing data [3].

Transparency about the software used as the essence of the scientific process is crucial to ensure reproducibility and to understand the provenance of individual research data and results. Even minor changes to the software might significantly influence the results of computational experiments [3].

The history of mathematical chemistry contains the contributions of many outstanding scientists, such as H. Weiner, A. T. Balaban, M. Randić, I. Gutman, N. Trinajstić, D. Bonchev, S. C. Basak, R. Carbó-Dorca, as well as many others [4–15].

Many scientific reviews have become available in this area—nevertheless, the most attractive ones consist of the quantitative features and characteristics of science and scientific research collected in the literature [16]. Nonetheless, success in mathematical chemistry in different fields, especially in drug design, has been and will continue to be on the verge of randomness and the danger of capital disappointments resulting from overly bold optimizations and globalization [17].

Biopharmaceutical companies have done everything possible in the last decade to globalize their capabilities. It is generally recognized that health information is a crucial external function that must continually focus on optimizing its capabilities to meet medical and even political challenges around the world [18]. However, the essential quality of the developed resources for a mathematical understanding of physical, chemical, and biochemical phenomena should be their open, general right of usage, that is, the data and results being accessible to a broad mass of users, from students to specialists working in other often distant fields, to apply QSPR/QSAR results.

1.2 QSPR/QSAR: Tools and Tasks

Developing new types of sweeteners, skin protection products, and cosmetics is costly; however, it requires economically suitable solutions. Drug research and development are even more complex, expensive, and time-consuming tasks requiring acceptable solutions. Quantitative structure–property/activity relationships (QSPRs/QSARs) are a popular approach to searching for answers to the above-listed functions and solutions to many others.

The attractiveness of QSPR/QSAR is caused by: (i) this is a concept of compact representation of complex physicochemical and biochemical phenomena; (ii) this is a more economical way of searching for appropriately defined aims substances in comparison with the experimental analysis; (iii) this is an additional way of knowing of the nature, in general, and (iv) this is a way to avoid or at least to reduce the use of animal tests drastically.

The wide variety of substances known now seems incomprehensible. This variety appears to be far from our understanding. However, applying computer technologies allows examining logically interacted parts of the above great list, at least fragmentally. To select substances for practical aims, detecting one quality (e.g., boiling point, heat capacity, toxicity, therapeutic effects, blood–brain barrier, and other properties) is not enough. Knowledge of the similarity and dissimilarity of the substance with others is necessary. The molecular structure is the basis for the comparison aimed at collecting the similarity and dissimilarity of different substances.

There are many manners to compare molecules—topology, i.e., connecting atoms, atoms composition, symmetry, and chirality. The 2D geometry represents these molecular features reliable enough. The diversity of molecular architecture rapidly increases in 3D space, starting from rotation conformers and finishing in supramolecular systems [19, 20]. The molecular biological systems demo incredible levels of simplicity and complexity. All living things depend on the ability of biomolecules to perform the functions of encoding and transmitting information, that is, to preserve and share various bio codes, including genetic ones [21–23].

The destruction of these molecular regulators leads to severe often-irreversible consequences in organisms. Predicting and controlling such damage is an ideal but hardly achievable goal of biochemistry and mathematical chemistry [24, 25]. Self-consistency and antagonism of molecular systems are also essential properties of biomolecules [26, 27]. Mathematical modeling of these phenomena is a complicated but essentially solvable problem. Moreover, for these purposes, there are pretty well-tested mathematical descriptions [28–30] and software available via the Internet [31–34].

The similarity and differences of molecules are considered. The results of such a comparison have been repeatedly described and found in numerous applications [31–34].

Of course, such comparisons are a very effective heuristic tool. However, it is no less exciting and promising to establish similarities and differences in various physicochemical and biochemical parameters from a heuristic point of view. Suppose the

comparison of molecules is carried out through configurations of atoms and bonds. In that case, the search for analogies and antagonism for arbitrary endpoints can be carried out based on comparisons of the corresponding structural fragments (alerts) that strongly or weakly affect the endpoints [35, 36]. However, the molecular structure is not very convenient for making numerous comparisons using a computer. The corresponding procedures are well implemented using molecular graphs or their representation utilizing special matrices. With multiple matrices representing different molecules and corresponding physicochemical or biochemical parameter values, molecular descriptors can be calculated using a single algorithm. The descriptors obtained by the manner can correlate with the endpoints of interest. Such correlations make it possible to build models calculated through linear regression equations obtained by the least squares method. These can be either models calculated using only one variable (descriptor) or models calculated using several or even many variables (descriptors).

This approach convinced many researchers that the high accuracy of such forecasts is possible since the correlation coefficients often showed very high values. Unfortunately, reality soon dispelled these high hopes. It turned out that a high correlation on the so-called training set was often accompanied by an extremely low correlation between the predicted and experimentally obtained values for physicochemical parameters (boiling points, melting points). Mainly, discouraging results were observed for biological activity (toxicity, drug efficacy). As a result, the term “chance correlation” appeared. Computer experiments have shown that, in principle, chance correlations can be recognized through the ratio of the number of descriptors involved in constructing the model and the number of molecules (substances) available for analysis. According to the Topliss-Costello rule [37], this ratio should be one to five (the ratio of the number of descriptors to the number of molecules). The problem of the dimensionality paradox and linear dependence also should be considered [38].

Despite these efforts devoted to improving the predictability of models, the QSAR practice has faced significant challenges, even if a group of several conceptual approaches to solving the same task is applied. Poor validation strategy is the most prevalent cause of the unsuitability of many QSAR models. The simple postulate “structurally those similar molecules should have similar biological properties” has also been seriously questioned and renamed the “QSAR paradox”. Such an occurrence is significant for the case of drugs since, as a rule, a drug should act on multiple targets rather than a single one. It increases the uncertainty of QSAR tasks and hence QSAR results related to drug discovery [38].

All these listed circumstances indicated that some reforms were needed in constructing and using QSPR/QSAR, both in theoretical and in practical terms. A contradiction or conflict often becomes a point of development, a transition to some new quality. Something similar happens from time to time in many, if not all, areas of the natural sciences. So, it happened with the QSPR/QSAR theory/practices.

The proclamation of the so-called Setúbal principles, which later became known as the “OECD principles”, can be considered a leap change in the paradigm of constructing “structure–property/activity” models.

1.3 Five OECD Principles

Gradually, the understanding of the necessity to check the statistical quality of a model for compounds unknown at the moment of building up the model did become the accepted principle.

The Organization for Economic Co-operation and Development (OECD) curates QSPR/QSAR studies. The agreed OECD principles are as follows: for real applying a QSAR model for regulatory purposes, it should be associated with the following information:

1. A defined endpoint;
2. An unambiguous algorithm;
3. A defined domain of applicability;
4. Appropriate measures of goodness-of-fit, robustness, and predictive potential;
5. A mechanistic interpretation, if possible.

Unfortunately, these principles are more legal than mathematical. But even in this capacity, they are instrumental.

1.4 Praxis of the QSPR/QSAR Development

Practice shows that in the field of QSPR/QSAR research, there are several paradigms (analytical comparisons of these may be helpful) for solving the problem of predicting the values of various endpoints (well-known to get truthful results if the comparison of two or more opinions is necessary). Most likely, the number of such paradigms will increase since none of the mentioned paradigms lacks both advantages and disadvantages. A brief overview of the paradigms used to build the QSPR/QSAR models follows, based on the diversity of molecular descriptors or algorithms for building models.

1.5 Molecular Descriptors are the Basis for the QSPR/QSAR

One might determine five construction levels according to the dimensionality of the spaces in which information is taken to calculate the molecular descriptor. A 0D descriptor is one for calculating which no information on the molecular structure is used (e.g., physicochemical property, solubility, or molecular weight). The 1D descriptor requires stoichiometric data for its calculation (e.g., the number of atoms or double/triple bonds). Then, 2D -descriptors are calculated according to molecular topology (configuration of atoms and bonds between atoms). The 3D

descriptors require data on molecular geometry (distances between atoms in three-dimensional space) for their calculation. And 4D descriptors are calculated by averaging all possible rotational conformers. Another version of 4D descriptors is the consideration of molecules in a relativistic space, where time (T) is considered a specific geometric component similar to the axes X, Y, and Z.

In practice, different methodologies apply all mentioned descriptors for QSPR/QSAR analysis, often without an attempt to elucidate—why this descriptor and no other one?

1.5.1 Principal Component Analysis

Principal component analysis (PCA) is probably the most popular multivariate statistical technique, and almost all scientific disciplines use it. PCA analyzes a data table representing observations described by several dependent variables, which are, in general, inter-correlated. Its goal is to extract the primary information from the data table and to express it as a set of new orthogonal variables called principal components. PCA also represents the pattern of similarity of the observations and the variables by displaying them as points in plots [39].

1.5.2 Multiple Linear Regressions

Multiple linear regression (MLR) is a statistical tool that uses independent variables to model dependent variable. The objective of MLR is to find a linear model of the property of interest according to the paradigm “Endpoint is a mathematical function of a group of descriptors” [40]. However, this leads to a vast labyrinth of possibilities; the number of options for combinations of descriptors grows exponentially with the growth of the number of available descriptors and the growth of the model dimensionality (three-, four-, ... n -dimension models).

1.5.3 Partial Least Squares

For structure–activity correlation, partial least squares (PLS) has many advantages over regression, including the ability to robustly handle more descriptor variables than compounds, non-orthogonal descriptors, and multiple biological results while providing more predictive accuracy and a much lower risk of facing the chance correlation. The significant limitations are a higher risk of overlooking “real” correlations and sensitivity to the relative scaling of the descriptor variables [41]. PLS

is the regression extension of PCA and is used for establishing QSARs. Judging by the number of mentions of PLS in the SCOPUS database, this method has a lot of supporters.

1.5.4 K-Nearest Neighbor Classification

K-nearest neighbors (KNN) is a nonparametric method used in the computational scheme of establishing the correlation “structure–property/activity” that specifies the class of each chemical based on K nearest of its neighbors (chemicals) from the training set. The class is equal to the type of the majority of the K neighbors of the tested chemical [42]. The molecular similarity is the basis of the approach.

1.5.5 Artificial Neural Network

Artificial neural networks are parallel computational devices consisting of groups of highly interconnected processing elements called neurons. Neural networks are characterized by topology, computational characteristics of their elements, and training rules. Traditional neural networks have neurons arranged in a series of layers. The first layer is termed the input layer, and each of its neurons receives information from the exterior, corresponding to one of the independent variables used as inputs. The last layer is the output layer; its neurons handle the output from the network. The layers of neurons between the input and output layers are called hidden layers. Each layer may make independent computations and pass the results to another layer. In feedforward neural networks, the connections among neurons are directed upwards, i.e., relationships are not allowed among the neurons of the same layer or the preceding layer. Networks where neurons are connected to themselves, with neurons in the same layer or neurons from a preceding layer, are termed feedback or recurrent networks. At a very simplified level, artificial neural networks mimic the way a biological brain organizes, stores, and processes information [43, 44].

The popularity of neural networks borders on complete trust in them; however, the emergence of hybrid approaches partially using neural networks indicates the possibility of improvements in “classical” neural networks [45].

1.5.6 Support Vector Machine

Support vector machine (SVM) is gaining popularity due to several attractive features and promising empirical performances. The primary aim of SVM is data classification, which is much easier and more applicable than artificial neural networks. Briefly, a classification task usually involves training and testing data which consists of some

data instances. Each instance in the training set contains one “target value” (class labels) and several “attributes” (features). The goal of SVM is to produce a model which predicts the target value of data instances in the test set, which is given only the attributes [46].

Thus, SVM has many advantages, but together with two disadvantages. These are the empirical nature of selecting descriptors and, as a rule, a large number of molecular features involved in a model [46].

1.5.7 Random Forest

The algorithm random forest is widely used in classification and regression, given that it has several features that make it suitable for QSAR/QSPR tasks. These include good predictive performance even when there are more variables than observations. The availability of measures of the ranging of descriptors and the ability to integrate a large number of simple models allow the possibility of reducing overtraining problems [47].

The main disadvantages are the possibility of the initial data influencing the predictive potential of the models, as well as the large amount of data required for the implementation of the models.

1.5.8 Monte Carlo Method

The main idea of the Monte Carlo method is to play a set of random changes in the simulation system, accompanied by quality control (evaluation) of the resulting models. The strength of this approach is its real objectivity (due to the random nature of all transformations). At the same time, the need to conduct many implementations/checks of these random modifications should be recognized as a weakness. Unfortunately, Monte Carlo methods cannot provide high accuracy in modeling anything, but they offer a comprehensive, absolutely random coverage of the phenomenon under study, that is, an analysis of even those situations that may seem illogical or unlikely; as a result, these possibilities escape from the attention of researchers. Here, so-called optimal 2D descriptors calculated by the Monte Carlo method for the defined endpoint are discussed [35, 36].

1.5.9 Data Curation

The curation of data selected for developing a model is a critically significant component of a QSPR/QSAR analysis. Previously, before the advent of computers and the Internet, an experiment was the primary data source. With the advent of computers,

the experiment has lost its privilege of being the sole source of data. It turned out that studying existing data (e.g., checking for consistency) can be a source of new data. Sometimes, the data should be averaged. Also, this includes procedures for identifying recent trends (new substances, new technologies) and new details in the behavior of complex biochemical objects. In other words, QSPR/QSAR analysis is possible only based on verified consistent data [48].

1.6 Reproducibility

A QSPR/QSAR loses significance if corresponding models are not reproducible in defined parametrization. At the same time, it must be taken into account that the ideal reproducibility of models with the appearance of new data (new substances, the establishment of additional factors affecting the physicochemical or biochemical behavior of molecular systems) is unattainable. However, the availability of reliable estimates of the dispersion of results indicates the reproducibility of predictive systems. QSPR/QSAR models should aim to meet a standard level of quality and be clearly described, ensuring their reproducibility [49].

In other words, each model should be checked up for a group of random splits into the training and validation sets.

1.6.1 *Applicability Domain*

One cannot apply a model if the domain of applicability of the model is not defined. The moment of determining the applicability domain is usually not considered. However, whether to determine the domain of applicability before building the model or the scope should be determined for the finished model nevertheless seems quite natural and quite important. From the practical point of view, the definition of the domain of applicability before building up a model appears more realistic. Appreciating the mechanisms is critical to determining the most likely applicability domain [50]. Four practical approaches for estimating the applicability domain in a multivariate space are applied: range, distance, geometrical, and probability density distribution [51].

1.6.2 *Model Validation*

Any QSPR/QSAR model becomes significant (suitable for practices) only after an appropriate assessment of the statistical quality of the model. Currently, there are no specific recommendations that suit everyone for assessing the predictive potential of models. The need for this kind of verification is noted, and the unreliability of the

existing criteria for the predictive potential is indicated, but the answer to the question “how can this be done?” remains in the realm of philosophy, that is, “practice is the reliability criterion of an approach”. In essence, this means that any approach is suitable for solving a problem if this approach has shown its ability to solve similar problems. Only models that have been validated externally can be considered reliable and applicable for external prediction: “validation” is the word that is constantly used but seldom defined [52].

A primitive and reliable model is possible, indicating the need for one or a few structural fragments to guarantee the desired effect. But in such a situation, nothing needs to be checked.

1.7 Recommendations for Building Robust QSPR/QSAR Models

QSAR is a collection of well-defined protocols and procedures that enable the definition of promising chemical collections [53, 54]. All QSPR/QSAR models are the result of computer experiments. One way or another, the identified molecular features line up in a series of factors contributing to increasing or decreasing the endpoint value. In some cases, unexpected analogies are observed between a computer experiment with numerical data and an actual physical experiment. For example, Fig. 1.1 shows the dependence of the number of poor predictions and the percentage of poor predictions for the validation set in group of models observed for different splits into the training and validation sets. One can see that 20% in the test set is preferable compared to the case where the test set contains 60% of the total set data. The graphic of the above dependence is similar to the graphic of the dependence of conductivity solutions of nanoparticles and their sizes [55]. Perhaps, this is a coincidence. Perhaps, there exists some invisible analogy between the computational process and the behavior of nanoparticles.

Therefore, it should be recognized that the formulation of a computer experiment, as well as any other experiment, requires the exclusion of the influence of the authors on this experiment. In other words, it is necessary to develop some standards that ensure the reproducibility of the results obtained, regardless of the conditions of a particular laboratory, well or poorly equipped, and irrespective of the personality of the researcher (only the latter must be conscientious enough so that the implementation of the instructions meets the necessary standards).

The traditional classical experiment with substances, energy, and information aims to formulate a question about nature and, secondly, to obtain an answer to this question. The computer experiment, in this sense, entirely coincides with the classical one despite being related mainly to information.

Below, some examples of applying the computational experiments based on the Monte Carlo method are discussed.

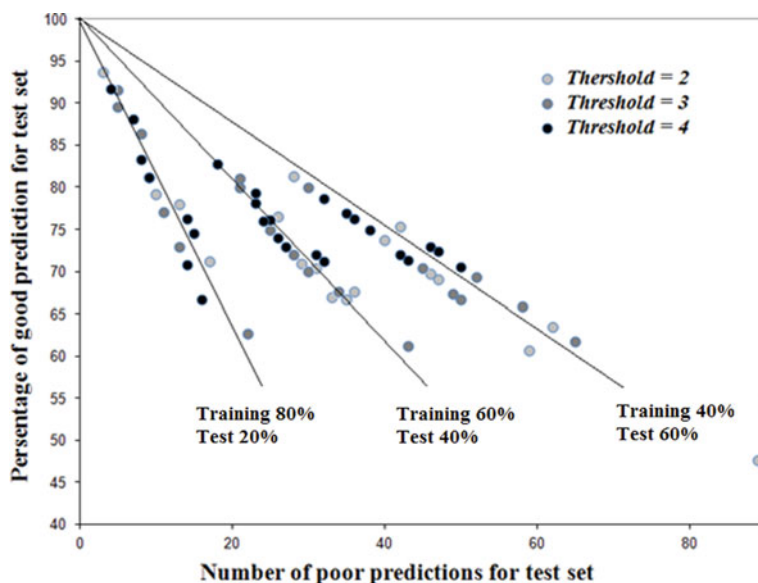


Fig. 1.1 Dependence of numbers of poor predictions and percentage of poor predictions for the validation set in group of models observed for different splits into the training and validation sets

1.8 Is It Possible to Obtain Correlations Suitable for QSPR/QSAR Using SMILES?

A positive answer to the first of the above questions was obtained from more than a hundred published works where the Monte Carlo technique (CORAL software, <http://www.insilico.eu/coral>) aimed to correlation weighting of molecular features to bring models of various physicochemical and biochemical endpoints collected in Table 1.1.

It should be noted that the development of optimal descriptors is possible not only based on SMILES, but it is also possible to develop optimal descriptors using both SMILES and the molecular graph (Table 1.2). The optimal descriptors of such categories were named hybrid ones [67–71]. Besides, optimal descriptors can be obtained from molecular graphs without using SMILES [72].

However, the practical use of a model should be in agreement with the research targets. SMILES and molecular graphs aim to represent the molecular structure. However, these representations are not identical. Can the superposition (hybrid) of these representations improve the quality of a model?

Table 1.1 Applying the Monte Carlo method to build up QSPR/QSAR models based on optimal descriptors

Endpoint	The statistical quality	Comments	References
Carcinogenic potency (pTD ₅₀)	$n = 170, R^2 = 0.628,$ RMSE = 0.87; $n = 61, R^2 = 0.758,$ RMSE = 0.602	The statistics for the training and validation set	[56]
The cellular uptake in PaCa2 cancer cells of nanoparticles	$n = 20, R^2 = 0.87,$ MAE = 0.15	The statistics for the validation set	[57]
Cytotoxicity for metal oxide nanoparticles	The statistical characteristics of these models are correlation coefficients 0.90–0.94 (training set) and 0.73–0.98 (validation set)	The average statistics on several splits	[58]
The mutagenic potential of multi-walled carbon nanotubes, pTA ₁₀₀	$n = 14, R^2 = 0.8087, Q^2 = 0.6975, s = 0.026, F = 51$ (training set); $n = 5, R^2 = 0.9453, s = 0.074$ (test set); $n = 5, R^2 = 0.8951, s = 0.052$ (validation set)	The approach checked up with three random splits	[59]
Cytotoxicity of different types of multi-walled carbon nanotubes to human lung cells	R^2 for internal validation datasets: 0.60–0.80; R^2_{pred} for external validation datasets: 0.81–0.88	Three random splits examined	[60]
Model for effective antidepressants, selective serotonin reuptake inhibitors	For the test sets of the four random splits, observed R^2 was 0.9459, 0.9249, 0.9473, and 0.9362	Four random splits examined	[61]
Aromatase inhibitors, a promising class of therapeutic anticancer agents (pIC ₅₀)	R^2 about 0.65(training set); R^2 about 0.68 (validation set)	Three random splits examined	[62]
Focal adhesion kinase inhibitors	The best statistical parameters $R^2 = 0.8398$ (validation set)	Four random splits examined	[63]

(continued)

Table 1.1 (continued)

Endpoint	The statistical quality	Comments	References
Acute toxicity of pesticides in rainbow trout (LC ₅₀)	training set: R^2 ranges 0.72–0.81, RMSE ranges 0.54–1.25; validation set R^2 ranges 0.74–0.84; and RMSE ranges 0.64–0.75	Three random splits examined	[64]
Biological activity of anti-diabetic drugs	$R^2 = 0.6837$ (training set); $R^2 = 0.8623$ (validation set)	Three random splits examined	[65]
Models for potential therapeutic SIRT1 for several diseases like cardiovascular, metabolic, and inflammatory disorders	$R^2 = 0.9524$ (training set), and $R^2 = 0.9058$ (test set)	Three random splits examined	[66]

Table 1.2 Applying the Monte Carlo method to build up QSPR/QSAR models based on hybrid optimal descriptors, which are calculated with SMILES, HSG, and GAO

Endpoint	The statistical quality	Comments	References
Biological activity of antihypertensive used in the treatment of hypertension, heart failure, and renal diseases	$R^2 = 0.8701$ (training set); $R^2 = 0.8430$ (test set)	Hybrid optimal descriptors are used, which are calculated with SMILES and HSG	[68]
Adsorption coefficients of aromatic compounds on multi-wall carbon nanotubes were studied	R^2 ranges 0.9463–0.8528 (training set); R^2 ranges 0.9573–0.8228 (validation set)	Hybrid optimal descriptors are used, which are calculated with SMILES and HSG	[69]
HIV-protease inhibitors (experimental inhibitory constant, Ki)	$n = 75$; $R^2 = 0.830$; RMSE = 0.489 (training set); $n = 15$; $R^2 = 0.915$; RMSE = 0.311 (validation set)	Hybrid optimal descriptors are used, which are calculated with SMILES and HSG	[70]
The prediction of binding affinities (pEC ₅₀)	The best statistical parameters $R^2 = 0.95$ (training set) $R^2 = 0.77$ (validation set)	Hybrid optimal descriptors are used, which are calculated with SMILES, HSG, and GAO	[67]
The prediction of binding affinities (pEC ₅₀)	The R^2 values of the three validation sets (splits 1 to 3) are 0.966, 0.921, and 0.886, respectively	Hybrid optimal descriptors are used, which are calculated with SMILES, HSG, and GAO	[71]

HSG Hydrogen suppressed graph; *GAO* Graph of atomic orbitals

1.9 The Main Quality of a Descriptor Is to Indicate the Differences Between Molecules

Most molecular descriptors correlate with molecular weight [73] and the length of the carbon chain or the carbon skeleton branching in organic molecules. An uncompleted list of molecular features that molecular descriptors should recognize includes the presence/absence of various rings, symmetry, and chirality [74]. In addition, desirable descriptors should “capture” the tendency of molecules to form intra- and intermolecular hydrogen bonds. The features mentioned above are quite interpretable. Descriptors that target correlation with the mentioned features are represented in the literature. However, the purpose of descriptors is to “capture” ultimately other abilities of molecules.

All models in the descriptor space might be wrong, but some are useful. How to prove that some model is valid?

Suppose a model’s construction is considered a particular event characterized by the values of statistical criteria. In that case, constructing a specific group of such models can be qualified as a group of random models. If the method is chosen adequately, then the statistical characteristics of these models should be more or less reproducible, albeit with some variance.

Having the statistical characteristics of groups of random in the above sense, models obtained by several methods, it is possible to compare the predictive potential of these methods. The method that gives the best statistical characteristics for external testing sets should be recognized as the most reliable for solving the problem.

To carry out the described computational experiments, (1) some set of compounds with experimental data on the considered endpoint is necessary; (2) a group of random distributions into a training set and a validation set; (3) a group of different methods for building up the model.

To confirm the above hypothesis, dataset on toxicity to Rainbow Trout of 309 pesticides (no mixtures) was taken in the literature [75]. Five random splits are calculated randomly using the CORAL software (<http://www.insilico.eu/coral>). These splits are random. The training sets are structured into three subgroups: active training set ($\approx 25\%$), passive training set ($\approx 25\%$), and calibration set ($\approx 25\%$). The external validation set also contains 25% of the total dataset. Three versions of hybrid optimal descriptors were used to develop a model for the above toxicity. The first hybrid descriptor calculated using SMILES and Morgan extended connectivity of the zero, first, and second order in HSG

$$\begin{aligned} \text{DCW}(T, N) = & \sum \text{CW}(S_k) + \sum \text{CW}(SS_k) + \sum \text{CW}(SSS_k) \\ & + \sum \text{CW}(\text{EC}0_k) + \sum \text{CW}(\text{EC}1_k) + \sum \text{CW}(\text{EC}2_k) \quad (1.1) \end{aligned}$$

The second hybrid descriptor calculated using SMILES and Morgan extended connectivity of the zero order in GAO

$$\text{DCW}(T, N) = \sum \text{CW}(S_k) + \sum \text{CW}(SS_k) + \sum \text{CW}(SSS_k) + \sum \text{CW}(\text{EC0}_k) \quad (1.2)$$

The second hybrid descriptor calculated using SMILES and Morgan extended connectivity of the zero and first order in GAO as follows:

$$\begin{aligned} \text{DCW}(T, N) = & \sum \text{CW}(S_k) + \sum \text{CW}(SS_k) + \sum \text{CW}(\text{EC0}_k) \\ & + \sum \text{CW}(\text{EC1}_k) \end{aligned} \quad (1.3)$$

In Eqs. 1.1–1.3, T is the threshold, i.e., the minimal frequency of SMILES attribute (S , SS , SSS) or Morgan's extended connectivity of zero, first, and second order (EC0_k , EC1_k , and EC2_k , respectively) in the active training set; N is the number of epochs of the Monte Carlo optimization applied to calculate the correlation weights (CWs) of the SMILES attributes and graph invariants.

The calculation of optimal descriptors needs the numerical data on the above correlation weights. Monte Carlo optimization is a tool to calculate those correlation weights. The target functions for the Monte Carlo optimization are the following:

$$\text{TF} = r_{\text{AT}} + r_{\text{PT}} - |r_{\text{AT}} - r_{\text{PT}}| \times 0.1 + \text{IIC} \times 0.5 \quad (1.4)$$

The r_{AT} and r_{PT} are correlation coefficients between the observed and predicted endpoints for the active and passive training sets. The IIC_C is the index of ideality of correlation [76, 77]. The IIC_C is calculated with data on the calibration set as follows:

$$\text{IIC}_C = r_C \frac{\min(-\text{MAE}_C, +\text{MAE}_C)}{\max(-\text{MAE}_C, +\text{MAE}_C)} \quad (1.5)$$

$$\min(x, y) = \begin{cases} x, & \text{if } x < y \\ y, & \text{otherwise} \end{cases} \quad (1.6)$$

$$\max(x, y) = \begin{cases} x, & \text{if } x > y \\ y, & \text{otherwise} \end{cases} \quad (1.7)$$

$$-\text{MAE}_C = \frac{1}{-N} \sum |\Delta_k|, \quad -N \text{ is the number of } \Delta_k < 0 \quad (1.8)$$

$$+\text{MAE}_C = \frac{1}{+N} \sum |\Delta_k|, \quad +N \text{ is the number of } \Delta_k \geq 0 \quad (1.9)$$

$$\Delta_k = \text{observed}_k - \text{calculated}_k \quad (1.10)$$

The observed and calculated are corresponding values of the endpoint.

Having the numerical data on the correlation weights, one can calculate the model via the equation

$$pLC_{50} = C_0 + C_1 \times DCW(T, N) \quad (1.11)$$

Tables 1.3, 1.4, and 1.5 contain the statistical characteristics of models calculated with the hybrid descriptors calculated using Eqs. 1.1–1.3, respectively.

Thus, the modeled properties' valid values should be sought in the average plus minus a variance format.

Figure 1.2 represents the average determination coefficient values observed for descriptors calculated with Eqs. 1.1–1.3 as well dispersion of these values. Thus, the comparison of the random QSAR models observed for different methods (Eqs. 1.1–1.3) indicated that the best method is the one observed for descriptor calculated with Eq. 1.2. In contrast, other methods are characterized by smaller determination coefficients for validation set and by more significant dispersion of this value.

Table 1.3 Statistical quality of the model is based on the optimal descriptor calculated with Eq. 1.1

Split	Set*	<i>n</i>	<i>R</i> ²	CCC	IIC	<i>Q</i> ²	RMSE	MAE	<i>F</i>
1	A	75	0.6695	0.8020	0.7553	0.6487	0.950	0.821	148
	P	80	0.7316	0.7504	0.7204	0.7171	1.06	0.906	213
	C	72	0.8069	0.8856	0.8981	0.7883	0.556	0.406	293
	V	82	0.7561				0.791	0.602	
2	A	74	0.7272	0.8421	0.8079	0.7099	0.856	0.729	192
	P	81	0.7141	0.8413	0.7562	0.7005	0.921	0.746	197
	C	77	0.7434	0.8616	0.8622	0.7305	0.682	0.554	217
	V	77	0.7785				0.680	0.533	
3	A	79	0.6703	0.8026	0.7588	0.6525	0.901	0.706	157
	P	76	0.7595	0.7509	0.7657	0.7495	1.06	0.942	234
	C	77	0.8886	0.9422	0.9425	0.8834	0.415	0.311	598
	V	77	0.7265				0.631	0.507	
4	A	78	0.6144	0.7611	0.7074	0.5946	0.976	0.855	121
	P	79	0.7157	0.6209	0.5627	0.7009	1.32	1.16	194
	C	76	0.8570	0.9233	0.9257	0.8504	0.380	0.288	443
	V	76	0.8270				0.515	0.431	
5	A	76	0.7702	0.8702	0.7492	0.7573	0.777	0.657	248
	P	77	0.7370	0.8474	0.7655	0.7211	0.895	0.753	210
	C	78	0.7734	0.8551	0.8794	0.7605	0.804	0.648	259
	V	78	0.6199				0.903	0.699	

*) Here and below, A, P, C, and V are active training, passive training, calibration, and validation sets, respectively

Table 1.4 Statistical quality of the model is based on the optimal descriptor calculated with Eq. 1.2

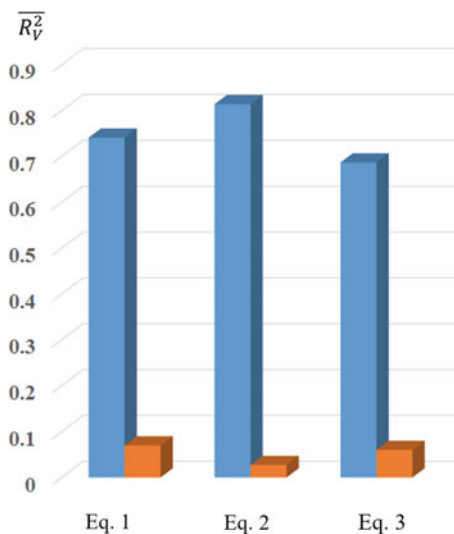
Split	Set	<i>n</i>	R^2	CCC	IIC	Q^2	RMSE	MAE	<i>F</i>
1	A	75	0.5006	0.6672	0.6191	0.4728	1.17	1.04	73
	P	80	0.7189	0.6233	0.8064	0.7056	1.18	1.07	199
	C	72	0.8489	0.9194	0.9213	0.8272	0.427	0.306	393
	V	82	0.8443				0.493	0.395	
2	A	74	0.7196	0.8369	0.6828	0.7006	0.868	0.730	185
	P	81	0.6099	0.7700	0.6277	0.5912	1.19	1.01	124
	C	77	0.8109	0.8978	0.9005	0.8017	0.601	0.499	322
	V	77	0.7966				0.676	0.496	
3	A	79	0.6764	0.8069	0.6885	0.6588	0.893	0.775	161
	P	76	0.6792	0.7455	0.7009	0.6653	1.11	0.956	157
	C	77	0.8637	0.9268	0.9292	0.8573	0.478	0.392	475
	V	77	0.7774				0.554	0.439	
4	A	78	0.5351	0.6972	0.6270	0.5056	1.07	0.977	87
	P	79	0.6353	0.6177	0.6677	0.6163	1.32	1.15	134
	C	76	0.8245	0.9010	0.9080	0.8162	0.430	0.339	348
	V	76	0.8470				0.495	0.397	
5	A	76	0.5481	0.7081	0.6320	0.7805	0.5213	1.09	0.957
	P	77	0.5098	0.7114	0.6547	0.8097	0.4850	1.27	1.13
	C	78	0.8346	0.9117	0.9135	0.8986	0.8239	0.536	0.447
	V	78	0.8067				0.528	0.420	

1.10 Significant Notes

- A QSPR/QSAR model is a random event (an unpleasant, ugly truth that cannot be ignored when building wrong, but perhaps useful, models).
- An approach should be estimated for a few different distributions into training and validation sets.
- The accurate measure of model robustness is likely to be the reproducibility of the statistical quality of the model across multiple splits into training and validation sets [78] rather than the high statistical quality of the model for a single split into training and validation sets.
- All published models built using CORAL software can be reproduced with an accuracy that users can measure by carrying out (repeated) corresponding computational experiments.

Table 1.5 Statistical quality of the model is based on the optimal descriptor calculated with Eq. 1.3

Split	Set	n	R^2	CCC	IIC	Q^2	RMSE	MAE	F
1	A	75	0.5207	0.6848	0.7026	0.4951	1.14	0.985	79
	P	80	0.7080	0.6379	0.7026	0.6916	1.20	1.08	189
	C	72	0.8230	0.9061	0.9072	0.8073	0.465	0.371	325
	V	82	0.6865				0.714	0.528	
2	A	74	0.6059	0.7546	0.6986	0.5789	1.03	0.823	111
	P	81	0.5314	0.7286	0.6403	0.5035	1.20	0.978	90
	C	77	0.7359	0.8448	0.8577	0.7221	0.661	0.521	209
	V	77	0.7151				0.705	0.558	
3	A	79	0.5649	0.7219	0.6621	0.5408	1.04	0.868	100
	P	76	0.6861	0.6892	0.7636	0.6651	1.17	0.983	162
	C	77	0.8155	0.8958	0.9030	0.8054	0.518	0.412	332
	V	77	0.6603				0.666	0.509	
4	A	78	0.4362	0.6074	0.5960	0.4006	1.18	1.06	59
	P	79	0.6106	0.5728	0.6648	0.5901	1.36	1.19	121
	C	76	0.7266	0.8139	0.8524	0.7118	0.557	0.416	197
	V	76	0.7758				0.588	0.452	
5	A	76	0.6432	0.7829	0.7609	0.6227	0.968	0.803	133
	P	77	0.5779	0.7566	0.7134	0.5552	1.22	1.02	103
	C	78	0.7659	0.8750	0.8751	0.7533	0.657	0.486	249
	V	78	0.5973				0.863	0.629	

Fig. 1.2 Comparison of the predictive potential of considered methods in building up models for toxicity of pesticides to Rainbow Trout

1.11 Conclusions

Estimating a physicochemical or biochemical parameter by QSPR/QSAR is a surrogate for a real experiment. However, the reproducibility of the results is necessary for assessing the QSPR/QSAR approach as successful. Despite the inconvenience of applying many criteria for the statistical quality of the model, if they are diverse in nature, they are the guarantors of the statistical reliability of the model and, therefore, the patrons of confidence in the used approach. The general philosophical significance of QSPR/QSAR lies in the satisfactory quality of the forecast of the phenomena under consideration and in the semantic load on obtaining and using QSPR/QSAR results. In other words, there must be harmony between the user and the logic of the program as a tool for solving the problem.

Acknowledgements AAT and APT are grateful to the project LIFE-CONCERT (LIFE17 GIE/IT/000461) for their support.

References

1. Wiener H (1947) *J Am Chem Soc* 69(1):17–20. <https://doi.org/10.1021/ja01193a005>
2. Wiener H (1947) *J Chem Phys* 15(10):766. <https://doi.org/10.1063/1.1746328>
3. Schindler D, Bensmann F, Dietze S, Krüger F (2022) *PeerJ Comput Sci* 8:e835. <https://doi.org/10.7717/PEERJ-CS.835>
4. Wiener H (1947) *J Am Chem Soc* 69(11):2636–2638. <https://doi.org/10.1021/ja01203a022>
5. Bonchev D, Trinajstić N (1977) *J Chem Phys* 67:4517–4533. <https://doi.org/10.1063/1.434593>
6. Balaban AT (1979) *Theor Chim Acta* 53(4):355–375. <https://doi.org/10.1007/BF00555695>
7. Bonchev D, Mekenjan Ov, Protić G, Trinajstić N (1979) *J Chromatogr A* 176(2):149–156. [https://doi.org/10.1016/S0021-9673\(00\)85645-9](https://doi.org/10.1016/S0021-9673(00)85645-9)
8. El-Basil S (1987) *Chem Phys Lett* 137(6):543–547. [https://doi.org/10.1016/0009-2614\(87\)80626-7](https://doi.org/10.1016/0009-2614(87)80626-7)
9. Gutman I, Miljković O, Caporossi G, Hansen P (1999) *Chem Phys Lett* 306(5–6):366–372. [https://doi.org/10.1016/S0009-2614\(99\)00472-8](https://doi.org/10.1016/S0009-2614(99)00472-8)
10. Gutman I, Araujo O, Morales DA (2000) *J Chem Inf Comput Sci* 40(3):593–598. <https://doi.org/10.1021/ci990095s>
11. Hansch C, Fujita T (1964) *J Am Chem Soc* 86(24):5710. <https://doi.org/10.1021/ja01078a623>
12. Dearden JC (2017) In: Leszczynski J (ed) *Challenges and advances in computational chemistry and physics*, vol 24, pp 57–88. https://doi.org/10.1007/978-3-319-56850-8_2
13. Doweiko AM (2008) *J Comput Aided Mol Des* 22(2):81–89. <https://doi.org/10.1007/s10822-007-9162-7>
14. Tóth G, Bodai Z, Héberger K (2013) *J Comput-Aided Mol Des* 27(10):837–844. <https://doi.org/10.1007/s10822-013-9680-4>
15. Weininger D (1988) *J Chem Inf Comput Sci* 28(1):31–36. <https://doi.org/10.1021/ci00057a005>
16. Mak K-K, Balijepalli MK, Pichika MR (2022) *Expert Opin Drug Discov* 17(1):79–92. <https://doi.org/10.1080/17460441.2022.1985108>
17. Segall MD, Beresford AP, Gola JMR, Hawksley D, Tarbit MH (2006) *Expert Opin Drug Metab Toxicol* 2(2):325–337. <https://doi.org/10.1517/17425255.2.2.325>
18. Giffin SA, Shah R, Soloff A, Vaysman AM, Oreper J, Gažo A, Gandhi P, Shah I, Malieckal T, Boulos D, Flowers T, Stevens CA, Rocco MS, Patel AS, Albano D (2019) *Ther Innov Regul Sci* 53(3):332–339. <https://doi.org/10.1177/2168479018779920>

19. Varnek A, Fourches D, Hoonakker F, Solov'ev VP (2005) *J Comput-Aided Mol Des* 19(9–10):693–703. <https://doi.org/10.1007/s10822-005-9008-0>
20. Thurston BA, Ferguson AL (2018) *Mol Simul* 44(11):930–945. <https://doi.org/10.1080/08927022.2018.1469754>
21. Rosandić M, Paar V (2022) *BioSystems* 218:104695. <https://doi.org/10.1016/j.biosystems.2022.104695>
22. Sanders J, Hoffmann SA, Green AP, Cai Y (2022) *Curr Opin Biotechnol* 75:102691. <https://doi.org/10.1016/j.copbio.2022.102691>
23. Kim S, Yi H, Kim YT, Lee HS (2022) *J Mol Biol* 434(8):167302. <https://doi.org/10.1016/j.jmb.2021.167302>
24. Kodama T, Ohtani H, Arakawa H, Ikai A (2005) *Ultramicroscopy* 105(1–4):189–195. <https://doi.org/10.1016/j.ultramic.2005.06.035>
25. Blaurock B, Hippeli S, Metz N, Elstner EF (1992) *Arch Toxicol* 66(10):681–687. <https://doi.org/10.1007/BF01972618>
26. Gagarin SG (1979) *J Struct Chem* 19(4):620–621. <https://doi.org/10.1007/BF00745694>
27. Fedorov VS (1975) *J Struct Chem* 15(5):794–797. <https://doi.org/10.1007/BF00747289>
28. Bongrand P (2022) *Curr Issues Mol Biol* 44(2):505–525. <https://doi.org/10.3390/cimb44020035>
29. Kampanarakis A, Farantos SC, Daskalakis V, Varotsis C (2012) *Chem Phys* 399:258–263. <https://doi.org/10.1016/j.chemphys.2011.07.031>
30. Anikin NA, Bugaenko VL, Kuzminskii MB, Mendkovich AS (2012) *Russ Chem Bull* 61(1):12–16. <https://doi.org/10.1007/s11172-012-0002-0>
31. Pramanik S, Roy K (2014) *Environ Sci Pollut Res* 21(4):2955–2965. <https://doi.org/10.1007/s11356-013-2247-z>
32. Coi A, Massarelli I, Murgia L, Saraceno M, Calderone V, Bianucci AM (2006) *Bioorg Med Chem* 14(9):3153–3159. <https://doi.org/10.1016/j.bmc.2005.12.030>
33. Katritzky AR, Kulshyn OV, Stoyanova-Slavova I, Dobchev DA, Kuanar M, Fara DC, Karelson M (2006) *Bioorg Med Chem* 14(7):2333–2357. <https://doi.org/10.1016/j.bmc.2005.11.015>
34. Karelson M, Maran U, Wang Y, Katritzky AR (1999) *Collect Czechoslov Chem Commun* 64(1):1551–1571. <https://doi.org/10.1135/cccc19991551>
35. Toropov AA, Toropova AP, Benfenati E, Salmona M (2018) *Toxicol Mech Methods* 28(5):321–327. <https://doi.org/10.1080/15376516.2017.1422579>
36. Toropova AP, Toropov AA, Begum S, Achary PGR (2018) *Curr Neuropharmacol* 16(6):769–785. <https://doi.org/10.2174/1570159X15666171016163951>
37. Topliss JG, Costello RJ (1972) *J Med Chem* 15:1066–1068. <https://doi.org/10.1021/jm00280a017>
38. Carbó-Dorca R (2021) *Pure Appl Chem* 93(10):1189–1196. <https://doi.org/10.1515/pac-2021-0112>
39. Abdi H, Williams LJ (2010) *Wiley Interdiscip Rev Comput Stat* 2(4):433–459. <https://doi.org/10.1002/wics.101>
40. Darnag R, Minaoui B, Fakir M (2017) *Arab J Chem* 10:S600–S608. <https://doi.org/10.1016/j.arabjc.2012.10.021>
41. Cramer RD III (1993) *Perspect Drug Discov Des* 1(2):269–278. <https://doi.org/10.1007/BF02174528>
42. Arian R, Hariri A, Mehridehnavi A, Fassihi A, Ghasemi F (2020) *Comput Biol Chem* 86:107269. <https://doi.org/10.1016/j.compbiolchem.2020.107269>
43. Niculescu SP (2003) *J Mol Struct: THEOCHEM* 622(1–2):71–83. [https://doi.org/10.1016/S0166-1280\(02\)00619-X](https://doi.org/10.1016/S0166-1280(02)00619-X)
44. Andrada MF, Vega-Hissi EG, Estrada MR, Garro Martinez JC (2015) *Chemom Intell Lab Syst* 143:122–129. <https://doi.org/10.1016/j.chemolab.2015.03.001>
45. Seifi A, Riahi-Madvar H (2019) *Environ Sci Pollut Res* 26(1):867–885. <https://doi.org/10.1007/s11356-018-3613-7>
46. Sun M, Zheng Y, Wei H, Chen J, Cai J, Jin M (2009) *QSAR Comb Sci* 28(3):312–324. <https://doi.org/10.1002/qsar.200860107>

47. Teixeira AL, Leal JP, Falcao AO (2013) *J Cheminform* 5(2):9. <https://doi.org/10.1186/1758-2946-5-9>
48. Ambure P, Cordeiro MNDS (2020) In: Roy K (ed) *Ecotoxicological QSARs. Methods in pharmacology and toxicology*. Humana, New York, NY, pp 97–109. https://doi.org/10.1007/978-1-0716-0150-1_5
49. Patel M, Chilton ML, Sartini A, Gibson L, Barber C, Covey-Crump L, Przybylak KR, Cronin MTD, Madden JC (2018) *J Chem Inf Model* 58(3):673–682. <https://doi.org/10.1021/acs.jcim.7b00523>
50. Schultz TW, Hewitt M, Netzeva TI, Cronin MTD (2007) *QSAR Comb Sci* 26(2):238–254. <https://doi.org/10.1002/qsar.200630020>
51. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T (2005) *ATLA Altern Lab Anim* 33(5):445–459. <https://doi.org/10.1177/026119290503300508>
52. Gramatica P (2007) *QSAR Comb Sci* 26(5):694–701. <https://doi.org/10.1002/qsar.200610151>
53. Tropsha A, Gramatica P, Gombar VK (2003) *QSAR Comb Sci* 22(1):69–77. <https://doi.org/10.1002/qsar.200390007>
54. Tropsha A (2010) *Mol Inform* 29(6–7):476–488. <https://doi.org/10.1002/minf.201000061>
55. Siao MD, Shen WC, Chen RS, Chang ZW, Shih MC, Chiu YP, Cheng C-M (2018) *Nat Commun* 9(1):1442. <https://doi.org/10.1038/s41467-018-03824-6>
56. Toropov AA, Toropova AP, Benfenati E (2010) *Eur J Med Chem* 45(9):3581–3587. <https://doi.org/10.1016/j.ejmech.2010.05.002>
57. Toropov AA, Toropova AP, Puzyn T, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2013) *Chemosphere* 92(1):31–37. <https://doi.org/10.1016/j.chemosphere.2013.03.012>
58. Toropova AP, Toropov AA, Rallo R, Leszczynska D, Leszczynski J (2015) *Ecotoxicol Environ Saf* 112:39–45. <https://doi.org/10.1016/j.ecoenv.2014.10.003>
59. Toropov AA, Toropova AP (2015) *Chemosphere* 124(1):40–46. <https://doi.org/10.1016/j.chemosphere.2014.10.067>
60. Trinh TX, Choi J-S, Jeon H, Byun H-G, Yoon T-H, Kim J (2018) *Chem Res Toxicol* 31(3):183–190. <https://doi.org/10.1021/acs.chemrestox.7b00303>
61. Veselinović AM, Milosavljević JB, Toropov AA, Nikolić GM (2013) *Eur J Pharm Sci* 48(3):532–541. <https://doi.org/10.1016/j.ejps.2012.12.021>
62. Worachartcheewan A, Mandi P, Prachayasittikul V, Toropova AP, Toropov AA, Nantasenamat C (2014) *Chemom Intell Lab Syst* 138:120–126. <https://doi.org/10.1016/j.chemolab.2014.07.017>
63. Kumar P, Kumar A, Sindhu J (2019) *SAR QSAR Environ Res* 30(2):63–80. <https://doi.org/10.1080/1062936X.2018.1564067>
64. Toropov AA, Toropova AP, Marzo M, Dorne JL, Georgiadis N, Benfenati E (2017) *Environ Toxicol Pharmacol* 53:158–163. <https://doi.org/10.1016/j.etap.2017.05.011>
65. Manisha, Chauhan S, Kumar P, Kumar A (2019) *SAR QSAR Environ Res* 30(3):145–159. <https://doi.org/10.1080/1062936X.2019.1568299>
66. Kumar A, Chauhan S (2017) *Drug Res* 67(3):156–162. <https://doi.org/10.1055/s-0042-119725>
67. Achary PGR (2014) *SAR QSAR Environ Res* 25(1):73–90. <https://doi.org/10.1080/1062936X.2013.842930>
68. Stoičkov V, Stojanović D, Tasić I, Šarić S, Radenković D, Babović P, Sokolović D, Veselinović AM (2018) *Struct Chem* 29(2):441–449. <https://doi.org/10.1007/s11224-017-1041-9>
69. Ahmadi S, Akbari A (2018) *SAR QSAR Environ Res* 29(11):895–909. <https://doi.org/10.1080/1062936X.2018.1526821>
70. Islam MA, Pillay TS (2016) *Chemom Intell Lab Syst* 153:67–74. <https://doi.org/10.1016/j.chemolab.2016.02.008>
71. Ahmadi S, Mehrabi M, Rezaei S, Mardafkan N (2019) *J Mol Struct* 1191:165–174. <https://doi.org/10.1016/j.molstruc.2019.04.103>
72. Toropov AA, Toropova AP (2002) *J Mol Struct: THEOCHEM* 581(1–3):11–15. [https://doi.org/10.1016/S0166-1280\(01\)00733-3](https://doi.org/10.1016/S0166-1280(01)00733-3)
73. Kurashov EA, Fedorova EV, Krylova JV, Mitrukova GG (2016) *Scientifica* 2016:1205680. <https://doi.org/10.1155/2016/1205680>

74. Crippen GM (2008) *Curr Comput-Aided Drug Des* 4(4):259–264. <https://doi.org/10.2174/157340908786786001>
75. Toropov AA, Toropova AP, Benfenati E (2020) *Aquat Toxicol* 227:105589. <https://doi.org/10.1016/j.aquatox.2020.105589>
76. Toropov AA, Toropova AP (2017) *Mutat Res-Genet Toxicol Environ Mutagen* 819:31–37. <https://doi.org/10.1016/j.mrgentox.2017.05.008>
77. Toropova AP, Toropov AA (2017) *Sci Total Environ* 586:466–472. <https://doi.org/10.1016/j.scitotenv.2017.01.198>
78. Majumdar S, Basak SC (2018) *Curr Comput-Aided Drug Des* 14(1):5–6. <https://doi.org/10.2174/157340991401180321112006>

Chapter 2

Molecular Descriptors in QSPR/QSAR Modeling



Shahin Ahmadi, Sepideh Ketabi, and Marjan Jebeli Javan

Abstract Molecular descriptors are mathematical representation of a molecule obtained by a well-specified algorithm applied to a defined molecular representation or a well-specified experimental procedure. The molecular descriptors as the core feature-independent parameters used to predict biological activity or molecular property of compounds in the quantitative structure property/activity relationship (QSPR/QSAR) models. Over the years, more than 5000 molecular descriptors have been introduced and calculated using different software. In this chapter, the main classes of theoretical molecular descriptors including 0D, 1D, 2D, 3D, and 4D-descriptors are described. The most significant progress over the last few years in chemometrics, cheminformatics, and bioinformatics has led to new strategies for finding new molecular descriptors. The different approaches for deriving molecular descriptors here reviewed, and some of the new important molecular descriptors and their applications are presented.

Keywords Molecular descriptors · QSAR · QSPR · Chemometrics · Chemoinformatic

Abbreviations

MoRSE	3D-Molecular Representation of Structures Based on Electron Diffraction
ACE	Angiotensin-Converting Enzymes
AFM	Atomic Force Microscopy
AZI	Augmented Zagreb Index
BET	Brunauer, Emmett, and Teller
CORAL	CORrelation And Logic

S. Ahmadi (✉) · S. Ketabi · M. Jebeli Javan
Department of Chemistry, Faculty of Pharmaceutical Chemistry, Tehran Medical Sciences,
Islamic Azad University, Tehran, Iran
e-mail: s.ahmadi@iautmu.ac.ir; ahmadi.chemometrics@gmail.com

DHFR	Dihydrofolate Reductase
DLS	Dynamic Light Scattering
EM	Electronic Microscopy
EDX	Energy Dispersive X-ray Spectrometry
ESEM	Environmental Scanning Electron Microscopy
FFF	Field Flow Filtration
FMO	Frontier Molecular Orbital Theory
HOMO	Highest Occupied Molecular Orbital
WW	Hyper-Wiener Index
ICPOES	Inductively Coupled Plasma Emission Spectroscopy
ICP-MS	Inductively Coupled Plasma Mass Spectrometry
LC	Liquid Chromatography
LUMO	Lowest Unoccupied Molecular Orbital
MW	Molecular Weight
MVC	Multivariate Characterization
PCA	Principal Component Analyses
PPs	Principal Properties
QSAR	Quantitative Structure–Activity Relationship
QSPR	Quantitative Structure–Property Relationship
SMILES	Simplified Molecular Input Line Entry System
TMACC	Topological Maximum Cross Correlation
TEM	Transmission Electron Microscopy

2.1 Introduction

2.1.1 History

The history of molecular descriptors as a feature vector for each compound is closely related to the concept of molecular structure [1]. The years between 1860 and 1880 were marked by a strong disagreement about the theory of molecular structure, which arose from studies on substances showing optical isomerism and Kekulé's (1867–1861) studies on the structure of benzene [2].

Today, many chemical, physical, and biological characteristics of compounds rely on the principle that these parameters are effects of its structural descriptors.

In 1868, Crum-Brown and Fraser [3] introduced first formulation about relationship between the bioactivity/property of a chemical (Φ) and its chemical constitution (C), as the following equation:

$$\Phi = f(C) \quad (2.1)$$

Based on this concept, many studies were conducted on the relationship of molecular descriptors to observed properties, including the relationship between the anesthetic power of various aliphatic alcohols with chain length of carbon and molecular weight [4], between the color of disubstituted benzenes with various ortho-, meta-, and para-orienting [5], and between the narcotic toxicity and solubility in water [6].

One of the most attractive quantitative structure–activity relationship (QSAR) approach is the Hammett equation [7]. In 1973, he showed a linear relationship between the rate constants of a series of methyl ester reactions with $N(\text{CH}_3)_3$ and the ionization equilibrium constants of the related carboxylic acids in aqueous solution at ambient temperature. The linear relationship between the ionization constant of the ester containing a substituent X in the meta (m) or para (p) orientation (K_X) and the ionization constant of the unsubstituted ester (K_H) is defined by the following formula:

$$\log\left(\frac{K_X}{K_H}\right) = \rho \cdot \sigma_X, \quad (2.2)$$

where σ_X is the constant of the substituent in m or p position is indicated by σ_m or σ_p , respectively. The absolute value of σ , which varies for each substituent, refers to the measure of the global electronic effect exerted on the reaction center by the presence of substituent X . The sign of σ is positive for electron-withdrawer and negative for electron-donor substituent. The electronic induction effect and the electronic resonance effect denote by σ_I and σ_R , respectively; the constant for the unsubstituted aromatic ring as a reference represented by σ_R^0 . Hammett's equation in this case defined by the following equation.

$$\log\left(\frac{K_X}{K_H}\right) = \rho_I \cdot \sigma_I + \rho_R \cdot \sigma_R^0 \quad (2.3)$$

2.1.2 QSPR/QSAR Modeling

In cheminformatics, a QSPR/QSAR model, either qualitative or quantitative, is a mathematical function that can be used to describe the connection between the molecular structures of a series of chemical compounds and their physicochemical properties/biological activities [8–14].

This field of knowledge assumes that the activity or property of a compound depends on its structural features, which affect its overall activities and properties [15–19].

Despite the formal differences between different methodologies, each QSPR/QSAR method is based on a QSPR/QSAR table that can be generalized as presented in Fig. 2.1 [20].

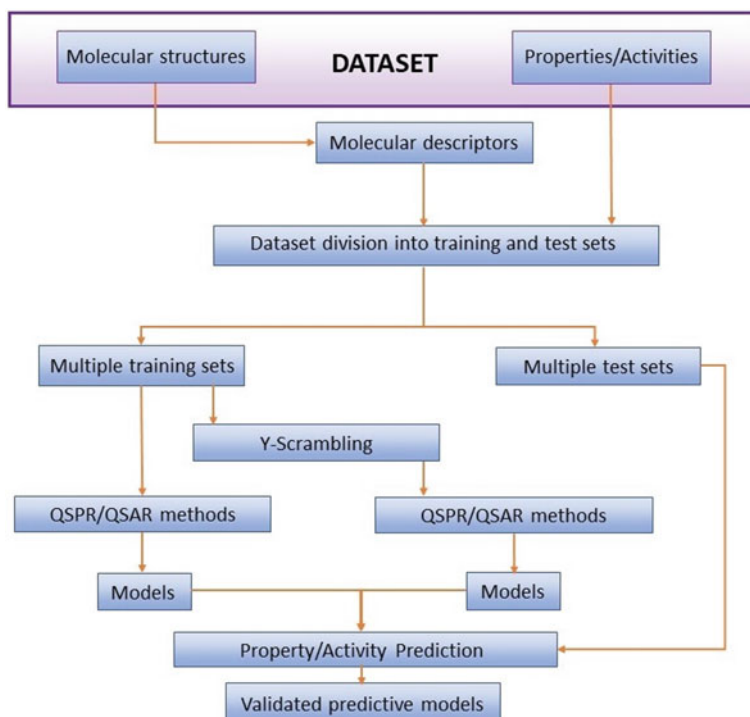


Fig. 2.1 Flowchart of the combinatorial QSAR methodology

The differences in various QSPR/QSAR studies can be explained in the following terms:

- Endpoint value
- Molecular descriptors
- Optimization algorithms.

Endpoint value as dependent variables can generally be of three types:

- Continuous

This endpoint is real values covering certain range, e.g., physicochemical properties of compounds such as boiling point and melting point. or IC_{50} values and binding constant.

- Categorical-related

This is classes of activities covering certain range of values, e.g., active and inactive compounds.

- Adjacent classes of metabolic stability

Adjacent classes of metabolic stability such as unstable, moderately stable, stable; and categorical-unrelated (i.e., classes of endpoints that do not relate to each other in any continuum, e.g., compounds that belong to different pharmacological categories, or compounds that are categorized as drugs vs. non-drugs).

Understanding this classification is indeed very important because the choice of descriptor types as well as modeling methods is often determined by the type of endpoints. Thus, in general the latter two types require classification modeling methods, whereas the former type of the target properties allows using linear regression modeling. Therefore, the latter two types require categorical modeling methods, generally while the former type of endpoint characteristics allows the use of linear regression modeling. Methods related to data analysis are called classification or continuous QSPR/QSAR.

2.1.3 Molecular Descriptors

Chemical descriptors as independent features in QSPR/QSAR modeling are usually classified into the following two types:

- Continuous

There are so many continuous descriptors such as molecular weight or many molecular connectivity indices.

- Categorical-related

The categorized descriptors such as counts of functional groups, binary descriptors indicating the presence or absence of a chemical functional group or an atom in a molecule.

2.1.3.1 Types of Molecular Descriptors

Molecular descriptors can be obtained from different representations of molecules. Knowing various types of descriptors is also critical for a fundamental understanding of QSPR/QSAR modeling because, as mentioned above, any modeling requires establishing a relationship between the chemical similarity of compounds and their target properties [21–24]. Chemical similarity is calculated in descriptor space using various similarity metrics [25]. For example, in the case of continuous molecular descriptors, the Euclidean distance in the descriptor space is an advisable choice of similarity metric, while in the case of binary descriptors metrics such as the Tanimoto coefficient or the Manhattan distance seem more appropriate.

The grade of the sufficiency of molecular structure samples differs from 0 to 4D demonstrations.

0D Descriptors

The 0D models contain the simplest molecule interpretation that does not hold any information about atom connections. Chemical formula, which organizes the atom types and their occurrences within a molecule, is independent of any information about the molecular structure. Therefore, molecular descriptors gained from the chemical formula stated as 0D descriptors. The most usual examples are atom type, number of atoms, molecular weight (MW), and any function of atomic properties.

1D Descriptors

Substructure list representation can be classified as a 1D description and contain of structural fragments of a molecule such as functional groups, bonds, rings, and substituents. Therefore, 1D descriptors do not involve a full information of molecular structure. These descriptors are inanimate to any conformation variation and, hence, do not recognize between isomers.

2D Descriptors

The 2D models include knowledge about the structure of the compound on the basis of its structural formula [26]. These patterns solely mirror the topology of the molecule. Such templates are highly common. The ability of such methods is that the topology model of the molecular structure includes information about the possible combinations of the molecule in virtual form.

Evaluation of the internal atomic arrangement of compounds is done by topological parameters [27]. They originated from the topological exhibition of molecules and can be measured as structure-manifest descriptors. These factors numerically code data related to molecular shape, size, branching, attendance of heteroatoms, and multifold bonds in numeric form. These topological parameters show the correlation of atoms by the characteristic of chemical bonds.

In modeling distinct biological, physicochemical, and pharmacokinetic properties, they have considerable performance. A topological display of the molecule is accessible as a molecular diagram. This diagram is defined in mathematical phrases as $G = (V, E)$, where V is a series of vertices corresponding to the atoms of the molecule and E is a series of elements that initiate a double connection between pairs of vertices.

These chemical diagrams illustrate a non-numerical figure of the molecular compound although a numeric interpretation of the diagram is crucial for computing topological parameters [28].

Some common 2D descriptors together with their description have been listed in the following.

Wiener (W) Index

The structure descriptor based on the classical molecular diagram is the Wiener index (W) which has become one of the most heavily applied descriptors in QSAR/QSPR approaches [29]. The descriptor is defined as the sum of edges on the shortest path in a chemical diagram.

Actually, the following equation denotes Wiener index $W(G)$ of the graph G (the graph G is a tree, T):

$$W(G) = \sum_{e \in E(G)} n_1(e|G)n_2(e|G) \quad (2.4)$$

$n_1(e|G)$ and $n_2(e|G)$ counts the vertices of G lying closer to the endpoints of the edge e than to its other endpoint

Hyper-Wiener Index (WW)

This index of a chemical tree T is defined as the sum of n_1n_2 products over all pairs of u vertices of T [30]. In fact, WW is the path number, and it is defined as the sum of the distances between any two atoms in the molecule, in terms of atom-atom bonds. Actually, WW can be calculated by multiplying the number of atoms on one side of any path by those on the other side, and the sum of these values for all paths. Wiener index is restricted to bonds and in Hyper-Wiener index bond is replaced with path.

Modified Wiener Index (W^*)

Bond contribution is determined by using the reciprocal of the number of atoms on each side of the bond [31].

Novel Wiener Index

It is obtained as an additive bond quantity, where the bond contribution is given as the product of the number of atoms close to each of the two points of each bond [32].

Connectivity Indices

It is structural invariant. Such indices are widely used in structure–property and structure–activity studies. These descriptors are on the basis of graph-theoretical constants that are presented to calculate the branching index of alkenes [33].

Kier and Hall extended these indices and intrinsic valence coupling indices to differentiate heteroatoms. Today, these phenomena have been optimized for a wide range of biological and physicochemical properties [34]. Randic [35] proposed some descriptors for topological indices: (i) they should be well-correlated with at least one feature; (ii) have structure commentary; (iii) be normal and self-determining; (iv) easily applied in a situational structure; (v) be free of empirical features; and (vi) be independent of other parameters.

Higher Order Connectivity

These indices are weight paths, where higher weight is given to terminal bonds and a lower weight to less exposed internal bonds [36].

Kier Shape

The descriptor defines shape indexes from molecular graphs. The shape of molecules is defined by the number of atoms and their bonding pattern which present in various orders [37].

Balaban Index

It is also one of the most distinctive molecular descriptors. Its value is independent of the molecular size or the number of rings [38].

Zagreb Indices

This descriptor is the first topological indices used for the total π -energy of conjugated molecules. The significant use of these indices is the distinction between the size of the molecules, flexibility, degree of branching, and entire shape [39].

Augmented Zagreb Index (AZI)

This index is based on the atom-bond connectivity (ABC index) used to obtain extreme values of AZI in chemical trees, and it can be used for upper and lower bonds' power of chemical trees [40].

Hosoya (Z)

It constructs QSAR/QSPAR models that describe the physical properties [41].

Modified Hosoya Index (Z*)

The frequency of occurrence of single CC bond in disjoint bond patterns is considered [42].

Autocorrelation Indices

This is a function of spatial separation and has particular advantageous for any QSAR/QSPAR study [43]

Szeged (SZ)

It is obtained as an additive bond quantity, where the bond contributions are given as the product of the number of atoms close to each of the two points of each bond [44].

Luckily, most of these parameters are identified in the topological descriptors. Therefore, they have been widely utilized in QSAR/QSPAR simulation to determine the structural resemblance or disparity of chemical compounds.

Topological Maximum Cross Correlation (TMACC)

These descriptors generated from atom properties determined by molecular topology based on concepts derived from autocorrelation descriptors. In 2007, Topological Maximum Cross Correlation (TMACC) was developed through atomic features characterized by molecular topology [45]. These parameters are based on meanings derived from coefficient descriptors. The ability to decode TMACC descriptors using QSAR simulation of angiotensin-converting enzymes (ACE) and dihydrofolate reductase (DHFR) inhibitors was demonstrated by Spowage et al. [46]. Altogether, TMACC revealed specific properties for C domain-selective ACE inhibition, which was an improvement on prior QSAR studies [46].

The physical and chemical features of a molecule that are evaluated by examining its 2D structure are physicochemical descriptors. These features play a main role in characterizing the drug condensation in the body. The convenient characteristics of a drug can enhance its effect and thus its market value.

Therefore, investigating these features of a drug not only contributes to the general plan of drug safety but also plays a significant role in drug detection collaboration by optimizing the selected compounds. Thus, it is necessary to pay attention to properties like solubility, permeability, and lipophilicity that can warrant optimal power, as well as to select the volunteer compounds with proper physicochemical properties.

The lipophilicity of a drug is related to its dependence on a lipophilic surrounding. It is an essential feature in the movement of drugs in the body, which includes intestinal absorption, membrane penetrance, protein linkage, and dispensation among multiple tissues [47].

Generally, a drug exhibits negligible chemical absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties in the presence of low lipophilicity [48]. Many pieces of research have been conducted on *in vitro* cellular permeance, which have demonstrated its connection to lipophilicity with other parameters, like molecular size, hydrophilicity, hydrogen bonds, and degree of ionization. These factors are recognized to have a considerable role in the intestinal absorption of a molecule. Molecular size is the main operative influencing biological activity like intestinal absorption.

Hydrogen bond donors and lipophilicity play considerable roles in predicting human intestinal permeability [49]. MW is associated with reduced permeability. Solubility in water plays a significant role in the distribution of drugs and their permeance through biological membranes, and their redeploy and sorption.

3D Descriptors

The 3D QSAR models [50–53] provide complete structural data including composition, topology, and steric form of the molecule for only one conformer. These patterns are the most common. Geometrical descriptors are computed from the 3D correlations of atoms in a given molecule. These parameters are in contrast to topological descriptors in terms of data and distinction power for similar chemical structures and molecular compounds [54].

In addition, they also contain data procured from atomic van der Waals regions and their participation on the molecular surface. In spite of their high data quantity, these parameters normally have drawbacks.

Geometrical descriptors need geometry optimization and, thus, the overhead to compute them. Thus, new data are available and can be extracted for flexible molecules that can have different molecular compositions. However, this propels the complexity that can enhance considerably. In addition, most of these parameters (grid-based descriptors) require arrangement rules to accomplish molecule abduction. Different groups of descriptors can be recognized using the set of geometric descriptors [54].

A diversity of 3D descriptors is accessible, some of them are:

3D-Molecular Representation of Structures Based on Electron Diffraction (MoRSE)

MoRSE descriptors have been shown to have good modeling power for various biological and physicochemical properties and can also be used to simulate infrared spectra [55].

Weighted Holistic Invariant Molecular (WHIM)

WHIM descriptors are applied to obtain related 3D data about molecular shape, size, symmetry and atom dispensation and have been utilized to model several physicochemical and toxicology properties. At the minimum, ten distinct sorts of WHIM parameters with distinct molecular characteristics have been expanded [54].

3D Autocohesion

Using the autocohesion function, these parameters are computed at individual spots on molecular surface. For a specific geometry and sensitive conformational change, they are unique and are constant to rototranslation [56].

GEometry, Topology, Atom-Weights Assembly (GETAWAY)

These parameters are on the basis of spatial coherence formula, which weights the atom to calculate van der Waals volume, atomic mass, and electronegativity alongside 3D data. According to data factors and the matrix operator, seven GETAWAY descriptors have been declared until now [54].

4D Descriptors

In 3D descriptors, the choice of the analyzed conformer is often random. The most adequate explanation of the molecular structure will be provided by 4D-QSAR patterns [57]. These models are similar to 3D models, but unlike them, structural data are discussed for a set of conformers (in essence, the fourth dimension), for a firm conformation.

Representation of molecular descriptors used in QSPR/QSAR modeling indicated in Fig. 2.2.

2.1.3.2 Molecular Descriptors' Resources

To get a considerable connection in QSAR studies, suitable descriptors must be used, whether they are empirical, theoretical, or derived from easily accessible experimental features of the molecules. Multiple descriptors mirror simple molecular features and thus can equip vision into the physicochemical characteristics of the property/activity under observation.

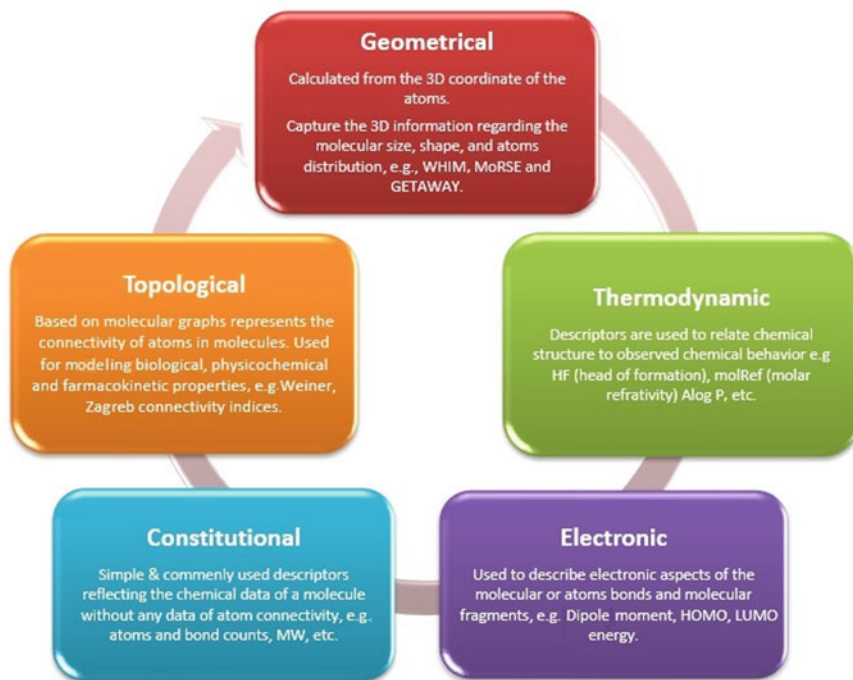


Fig. 2.2 Representation of molecular descriptors used in QSPR/QSAR modeling

Quantum Chemical Descriptors. Quantum chemical computations are an important source of new molecular descriptors that can actually represent all electronic and geometrical properties of molecules and their interactions.

Quantum chemical and molecular modeling techniques provide the description of a large number of molecular and local values that determine the shape, reactivity, and binding characteristics of an entire molecule in addition to its molecular pieces and substituents.

In the last years, quantum chemical parameters have been significant in QSAR models helping researchers illustrate the biological activities and toxicity mechanisms of various chemicals. In the past decades, semiempirical calculations were the prior ways to generate descriptors owing to the restrictions of the software and applied systems. Recent advances in computational hardware and the expansion of effective algorithms have helped to expand molecular quantum mechanical computations. In particular, the parameters derived from density functional theory (DFT) and hybrid density functional calculations (mPW1PW91) have excellent potential through their better accuracy in contrast to the semiempirical procedure and have good efficiency to fit into the geometrical, electrostatic, and orbital energy calculations [58–61].

Since the context of large discrete physical data is encoded in a large number of theoretical descriptors, their usage in the scheme of instruction sets in QSAR studies offers two significant priorities: (a) molecules, their diverse parts, and their

substitutions; can be instantly identified based on their molecular structure, and (b) the presented mechanism of action can be straight considered for the chemical reaction of the studied compounds [62]. As a result, the derived QSAR models contain data on the essence of the intermolecular interactions imported in specifying the biological or other properties of the investigated compounds. The most commonly used quantum chemical descriptors can be classified as follows:

Geometry Descriptors. The bond lengths, angles, and molecular dihedrals of the root segment should be the same for all molecules in the series.

Atomic Charges. In accordance with the classical theory of chemistry, all chemical interactions are either orbital (covalent) or electrostatic (polar) in nature. The electric charges in the molecule are clearly the order of the electrostatic interactions. Indeed, local electron density or charges have been shown to be momentous in a large number of physicochemical properties and chemical reactions of structures. Therefore, charge-based descriptors have been broadly utilized as indicators of chemical reactivity or as a measure of fragile intermolecular interactions. Numerous quantum chemical descriptors are derived from partial charge. Partial atomic charges are known as indicators of static chemical reactivity [63]. The computed σ - and π -electron densities on a specific atom determine the feasible direction of the chemical interactions and, hence, are often discussed as indices of directional reactivity. Unlike the total electron density, specific charges on atoms are observed as indicators of non-directional reactivity. Several sums of absolute or squared values of partial charges have also been used to characterize intermolecular interactions, e.g., solute–solvent interactions [64–66].

Molecular Orbital Energies. Highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies are very universal quantum chemical descriptors. It has been displayed [67] that these orbitals play an important role in controlling various chemical reactions and specifying electronic band gaps in solids. They are also in charge of the formation of several charge transfer complexes [63, 68]. Based on the frontier molecular orbital theory (FMO) of chemical reactivity, the organization of a transition state is owing to the interaction between the frontier orbitals (HOMO and LUMO) of the reacting fragments [69]. Therefore, the behavior of frontier molecular orbitals is distinct from others based on the general origins controlling the character of chemical reactions [69]. The HOMO energy is straightly connected to the ionization potential and characterizes the ability of the molecule to attack by electrophiles. The LUMO energy is straightly connected to the electron affinity and determines the readiness of the molecule against nucleophile attack. Both the HOMO and the LUMO energies are essential in radical reactions [70, 71]. The meaning of soft and hard nucleophiles and electrophiles is also connected to the relative energy of the HOMO/LUMO orbitals.

Soft nucleophiles have high-energy HOMOs. Hard nucleophiles have low-energy HOMOs. Soft electrophiles have low-energy LUMO, and hard electrophiles have high-energy LUMOs[72]. The HOMO–LUMO gap, i.e., the energy difference between HOMO and LUMO, is a major stability indicator [73].

$$E_{\text{gap}} = E_{\text{LUMO}} - E_{\text{HOMO}} \quad (2.5)$$

A large HOMO–LUMO gap indicates high resistance for the molecule by definition its less reactivity in chemical reactions [67]. The HOMO–LUMO gap has also been utilized as an estimate of the lowest stimulation energy of the molecule. However, this definition ignores electronic restructuring in the excited state and hence may mostly make incorrect theoretical results. The meaning of activation hardness (η) and softness (S) is also determined based on the HOMO–LUMO energy gap.

$$\eta = \frac{(E_{\text{LUMO}} - E_{\text{HOMO}})}{2} \quad (2.6)$$

$$S = \frac{1}{2\eta} \quad (2.7)$$

Activation hardness determines the rate of reaction at various sites of the molecule and is therefore related to anticipating direction effects [67]. The qualitative description of hardness is intimately connected to polarizability, as a reduction in the energy gap normally results in an easier polarization of the molecule [74].

Frontier Orbital Densities. Frontier orbital electron densities on atoms provide an effective alternative or accurate description of donor–acceptor interactions [71, 75]. Due to the theory of frontier electron reactivity, most chemical reactions happen in the location and direction where the overlap of the HOMO and LUMO of the respective reactants can be maximized [69].

In the matter of a donor molecule, both ionization potential (IE) and HOMO density (electrophilic electron density, f_r^E) are necessary to charge transfer:

$$f_r^E = \sum (C_{\text{HOMO},n})^2; \quad C_{\text{HOMO},n} \text{ are atomic orbital factors in HOMO} \quad (2.8)$$

$$\text{IE} = -E_{\text{HOMO}} \quad (2.9)$$

and in the terms of an acceptor molecule, LUMO density (nucleophilic electron density, f_r^N) and electron affinity (EA) are critical [63].

$$f_r^N = \sum (C_{\text{LUMO},n})^2; \quad C_{\text{LUMO},n} \text{ are atomic orbital factors in LUMO} \quad (2.10)$$

$$\text{EA} = -E_{\text{LUMO}} \quad (2.11)$$

These descriptors have been applied in QSAR studies to characterize drug–receptor interaction sites. By comparing the relativities of different molecules, the frontier electron density should be normalized by the energy of the frontier molecular

orbitals, and hence molecules with lower ionization potentials are predicted to be more reactive as nucleophiles. Absolute electronegativity index (χ), electron affinity (ω), and electron charge transfer (ΔN) are also determined based on ionization potential and electron affinity:

$$\chi = \frac{(I + A)}{2} \quad \text{absolute electronegativity} \quad (2.12)$$

$$\omega = \frac{\mu^2}{2\eta} \quad \text{electrophilicity index} \quad (2.13)$$

$$\Delta N = \frac{(\mu_B - \mu_A)}{2(\eta_A + \eta_B)} \quad \text{electron charge transfer} \quad (2.14)$$

Molecular Polarizability. The polarization of a molecule by an external electric [76] area is given by the potential tensors of order n of the molecular mass. The first-order term is used as polarizability (α):

$$\alpha = \frac{1}{3}(\alpha_{xx} + \alpha_{yy} + \alpha_{zz}) \quad (2.15)$$

The second-order term is mentioned in the first hyperpolarizability, etc. Therefore, the most considerable characteristic of molecular polarizability is binding to the molecular bulk or molar volume [73]. Polarizability values have been demonstrated to depend on hydrophobicity and other biological activities [77–79]. In addition, the electronic polarizability of the molecules contributes to the typical parameters of electrophilic super-delocalizability [80]. The first-order polarizability tensor includes data about feasible inductive interactions in the molecule [70, 73, 81, 82]. The total anisotropy of the polarizability (second-order term) determines the properties of a molecule as an electron acceptor:

$$\beta^2 = \frac{1}{2}[(\alpha_{xx} - \alpha_{yy})^2 + (\alpha_{yy} - \alpha_{zz})^2 + (\alpha_{zz} - \alpha_{xx})^2] \quad (2.16)$$

Dipole Moment and Polarity Indices. The polarity of a molecule is essential for several physicochemical properties. A large number of descriptors have been suggested to estimate the polarity effects. For instance, molecular polarity counts for chromatographic retention in a polar static phase [65, 83]. The dipole moment (μ) is the most obvious and is often used to explain the polarity of the molecule [64, 65, 70, 81, 84]. Difference between net charges on atoms (Δ) [68, 84], and topological electronic index (T_E) [68].

$$T_E = \sum_{i,j,i \neq j} \frac{|q_i - q_j|}{r_{ij}^2} \quad (2.17)$$

The quadrupole moment tensor can also be applied as an index to characterize probable electrostatic interactions. However, such tensors belong to the selection of the coordinate system and thus the direction of the molecular root section must be the same for all molecules in the series [70].

Energy. The total energy computed by quantum mechanical methods has been presented as a good descriptor in several cases [64, 68, 85, 86].

In addition, thermodynamic parameters contain entropy (S°), internal energy (Eth), constant-enthalpy (H°), free energy (G°), zero-point vibrational energy (ZPE), and volume heat capacity (CV°) can be computed from frequency quantum mechanical calculations. Reaction enthalpy (ΔH), entropy (ΔS), and free energy (ΔG) can be calculated by the difference in heats of formation, entropy, and free energies of formation between reactants and products or between conjugate forms [87, 88]. The protonation energy, described as the difference between the total energy of the protonated and neutral forms of the molecule, can be discussed as a good scale of the power of hydrogen bonds (the higher the energy, the stronger the bond) and can be used to specify the correct position of the most desirable hydrogen bond acceptor [89].

The others. The descriptors considered above form the bulk of quantum chemical descriptors effectively used in QSAR/QSPR studies. Other descriptors have also been designed but do not fall into the categories mentioned above, such as frequency and NMR chemical shifts.

2.1.3.3 Empirical and Experimental Descriptors

Quantum chemical and molecular modeling techniques allow the description of many molecular and local values that determine the reactivity, binding features, and shape of a molecule in addition to molecular moieties and substituents. A principled combination of theoretical molecular descriptors with both empirical Hammett's substituent constants (σ_m and σ_p) [90, 91], Swain–Lupton's field and resonance constants (F and R) [92], hydrophobic constant (Π) [92], Taft's steric parameter (E_s) [92], Verloop's steric parameters [90, 91], etc., and experimental descriptors (substituent-induced chemical shifts, molecular weight and molecular refractivity (MR) [92]) are available. Table 2.1 shows the list of empirical and experimental descriptors.

The mentioned substituent descriptors can be categorized pursuant to three main cluster groups: (a) descriptors that capture the effects of the substituent on the aromatic ring (electronic charges on the ring carbon atoms, resonance and field substituent constants, and substituent-induced chemical shifts); (b) descriptors characterizing the properties of the majority of substituents (Verloop's steric parameters and the molecular refractivity) are clustered with theoretical descriptors describing the polarizability properties of the substituents, molecular polarizability anisotropy, dispersion interaction terms (IP*ANIS, IP* $\Sigma\Pi_{\text{mol}}$) and electrophilic super-delocalizability of the substituent.

Table 2.1 List of empirical and experimental descriptors

Descriptor	Definition	References
σ_x	Taft's substituent electronegativity effect parameter	[93]
σ_α	Taft's substituent polarizability effect parameter	[93]
σ_f	Taft's substituent field effect parameter	[93]
σ_r	Taft's substituent resonance effect parameter	[93]
C_0	^{13}C substituent chemical shift on the ortho-carbon atom	[94]
C_i	^{13}C substituent chemical shift on the ipso-carbon atom	[94]
C_m	^{13}C substituent chemical shift on the meta-carbon atom	[94]
C_p	^{13}C substituent chemical shift on the para-carbon atom	[94]
σ_m	Hammett's substituent constant for the meta position	[90, 91]
σ_p	Hammett's substituent constant for the para position	[90, 91]
F	Swain–Lupton's field constant	[92]
R	Swain–Lupton's resonance constant	[92]
Π	Π hydrophobic constant	[92]
MR	Molecular refractivity	[92]
E_s	Taft's steric parameter	[92]
H_a	Number of hydrogen bonds that the substituent can accept	[95]
H_d	Number of hydrogen bonds that the substituent can donate	[95]
L	Verloop multidimensional steric parameter	[90, 91]
B_1	Verloop multidimensional steric parameter	[90, 91]
B_2	Verloop multidimensional steric parameter	[90, 91]
B_3	Verloop multidimensional steric parameter	[90, 91]
B_4	Verloop multidimensional steric parameter	[90, 91]
μ_{ar}	Lien's group dipole moment for aromatic substituent	[22]
λ_{ar}	Testa's lipophobic constant for aromatic substituent	[95]

IP = ionization potential derived from the AM1 wave function.

ANIS = anisotropy of the molecular polarizability.

IP*ANIS = product of the molecular ionization potential and the anisotropy of the molecular polarizability.

IP* $\Sigma\Pi_{\text{mol}}$ = product of the molecular ionization potential and the sum of the self-atom polarizability over all the atoms of the molecule.

$\Sigma\Pi_{\text{XX}}$ = sum of the self-atom polarizability values of the substituent atoms.

$\Sigma\Pi_{\text{mol}}$ = sum of the self-atom polarizability over all the atoms of the molecule.

ΣS_X^H = sum of the electrophilic super-delocalizability on the substituent atoms.

$\Sigma S_{E.X}$ = sum of the electrophilic super-delocalizability (computed over all the occupied molecular orbitals) on the substituent atoms.

$\Sigma S_{N.X}$ = sum of the nucleophilic super-delocalizability (computed over all the unoccupied molecular orbitals) on the substituent atoms.

The hydrophobic parameter Π is near to this cluster and to the solvent hydrophobic available surface of the substituent and the electrophilic super-delocalizability with the polarizability of the benzene ring; (c) molecular dipole moments and their experimental and theoretical substituents and their square.

(a) Hammett substituent constants, substituent-induced chemical shifts, and Taft and Lupton's resonance constants are mapped by the first component, the major contribution of which is the electronic charges of the carbon atoms of the benzene ring, the super-electrophilic mobility of the benzene ring and the energy of frontier molecular orbitals; (b) Verloop steric descriptors and the molecular refraction along with substituent van der Waals volumes and molecular weight are mapped by the second principal component, which includes theoretical parameters described as polarizability ($\Sigma\Pi_{XX}$, ANIS, $\Sigma\Pi_{\text{mol}}$), dispersion forces ($\text{IP}^*\Sigma\Pi_{\text{mol}}$, IP^*ANIS), and substituent reactivity indices (ΣS_X^H , $\Sigma S_{E,X}$, and $\Sigma S_{N,X}$). These recent cases perhaps indicate the portion of the molecular orbital development to molecular shape; (c) the third component models the lipophobic descriptor λ_{ar} and the lipophilic descriptor Π . The parameters that collaborate to this part are the dipole moments (consisting of the group dipole moment, μ_{ar}) and their square terms, the solvent available surfaces of the substituent, the energy difference between the HOMO and the LUMO (GAP), the Π -symmetry component of the electronic charges and the polarizability of the ring.

However, λ_{ar} and Π are not solely modeled by this section, as they also contribute significantly to the first and the third components, respectively. This suggests that more than one type of substituent effect specifies the values of these parameters. The same result is for the steric descriptors E_s modeled both by the first and the second components. These findings are similar to other research aimed at modeling Π [96] and E_s [97] and support the intricate character of these empirical parameters.

Empirical scales called principal properties (PPs) which define the physicochemical features of twenty naturally encoded amino acids were recently developed by Sjoström and Wold [98].

Sjoström et al. applied the PPs in the same way to categorize several types of signal peptides of different lengths [99]. Carlson and co-workers have reported principal component analyses (PCA) of multivariate characterization (MVC) characterize PPs, the physicochemical properties of organic solvents [100], Lewis acids in organic synthesis [101], amines in the Willgerodt Kindler reaction [102], and aldehyde/ketones [103].

These PPs are now heavily used in their laboratory to explore the realm and limits of new organic reactions. PPs of amino acids may be suitable for instance for screening of peptides [104]. The expansion of PPs for many aromatic substituents for subsequent uses has been the aim of researchers, and unfortunately, it is very difficult to find experimental information evaluated in a coordinated manner on a large number of substituents. Therefore, they should use the next best kind of data, famous and broadly used physicochemical parameters that are accessible for a large number of substituents.

The empirical parameter used to characterize a class of monosubstituted benzenes were Π , MR, σ_m , σ_p [92, 105], and the Verloop descriptors L and B_1 – B_4 [106]. The Verloop parameters B_1 – B_4 , derived from STERIMOL calculations, are normally listed in order of magnitude improvement. Researchers attempt to choose the variables to define steric bulk (MR), hydrophobicity (Π), the shape of each substituent (Verloop parameters), and electronic properties (sigmas).

In this case, they knew that there are three groups of variables: hydrophobicity/bulk, electronic, and size.

From the numeric amounts of the loadings, it is shown that the first component is significantly connected to the steric bulk and hydrophobicity because the length, molecular refractivity, and Π have the largest contributions. The second component is dominated by the two electronic descriptors, σ_m and σ_p , while the third component is again mainly hydrophobicity (Π) but also shape since L and B_1 – B_4 (Verloop parameters) [106] have relatively large contributions.

Since biological sieving of chemical substances is both expensive and time-consuming, it is essential to expand an instrument for the statistical design of the compounds in a filtering experiment. The main features are heavily appropriate for this purpose because they are few and orthogonal.

2.2 Descriptors for Nano-QSPR/QSAR

Over the past few decades, nano-based technology has become one of the top research areas in all fields of science and technology. A wide variety of consumer products are at the nanoscale, typically defined by all species having at least one diameter of 100 nm or less. Currently, nanotechnology has integrated various fields including biomedicine, pharmaceutical industry, food industry, environmental protection, solar batteries, energy, information and communication, heavy industry, consumer goods, and so on. However, it seems that we are only at the beginning of the “nano-industrial revolution.” Because of the unique electrical as well as optical, magnetic, thermal, and chemical properties of nanomaterials, the range of their possible applications is likely to expand rapidly.

Some recent papers report obvious evident toxicity of selected nanoparticles and highlight potential risk associated with the development of nano-engineering. Currently, there are many gaps in nanomaterial data. Predictive nano-QSAR/QSPR is one of the most promising methods used by chem informaticians to extrapolate the activity/property of nanomaterials. We believe that some of the missing data that are crucial for environmental risk assessment can be obtained using computational chemistry, saving the time and cost of conducting experiments. It is worth noting that the nano-QSPR/QSAR approach should be employed to predict not only activity responses (e.g., toxicity) but also many important physicochemical properties (e.g., water solubility, n-octanol/water partition coefficient, vapor pressure). These physicochemical properties affect the absorption, distribution, and metabolism of the compound in the organism, as well as environmental transport and the fate.

In nano-QSPR/QSAR modeling, one of the important parameters for building a validate model is suitable descriptors. In general, there are more than 5000 different descriptors for the characterization of molecular structure from zero to four dimensional (0D–4D). Only a few of traditional descriptors can characterize nanostructures. There are some reports that [107, 108] the existing descriptors are not enough to express the specific physical and chemical properties of nanoparticles. Therefore, new and more suitable types of descriptors for characterizing of nanoparticles should be developed.

Even though the computational features used for QSPR/QSAR modeling, experimentally derived features may also be employed as descriptors for nano-QSARs development (Fig. 2.3). The experimental descriptors seem to be especially useful for expressing size distribution, aggregation mode, shape, porosity, and surface disorder. Moreover, the combination of experimental results with a numerical approach can be used to define a new descriptor. For instance, images obtained by scanning electron microscopy (SEM), transmission electron microscopy (TEM), or atomic force microscopy (AFM) might be processed with new chemometric methods of image analysis. This means that first a series of pictures of different particles of a nanostructure should be taken. Then, the images must be numerically averaged and converted into a matrix containing numerical values that correspond to each pixel's grayscale intensity or red, green, and blue (RGB) color value. The other descriptors can be produced based on the matrix (i.e., the shape descriptor can be obtained as the sum of the nonzero elements in the matrix; the porosity as the sum of the relative differences between each pixel and its "neighbors," etc.) [109].

Undoubtedly, proper characterization of nanoparticle structure is currently one of the most challenging tasks in nano-QSAR. Although more than five thousand QSAR descriptors have been defined until now, they may be insufficient to express the supramolecular phenomena governing the unusual activity/property of nanomaterials. Consequently, much more effort is needed in this area.

2.3 SMILES and Quasi-SMILES Descriptors

The CORrelation And Logic (CORAL) software (<http://www.insilico.eu/coral/>) was developed by Alla Toropova and Andrey Toropov used to build up QSPR/QSAR models using Simplified Molecular Input Line Entry System (SMILES) [61, 111–116] and quasi-SMILES descriptors. SMILES is a chemical notation system designed by Weininger et al. [117, 118]. According to the principles of molecular graph theory, SMILES uses a very small, natural grammar to specify precise structural features. The SMILES symbol system is also suitable for high-speed machine processing [119, 120].

Over the last two decades, there have been numerous reports on the QSAR/QSPR modeling of nanomaterials and other compounds using CORAL software. This approach provides simple representation of molecular structures. There are defined equivalences between the representation of molecular structure using diagrams and

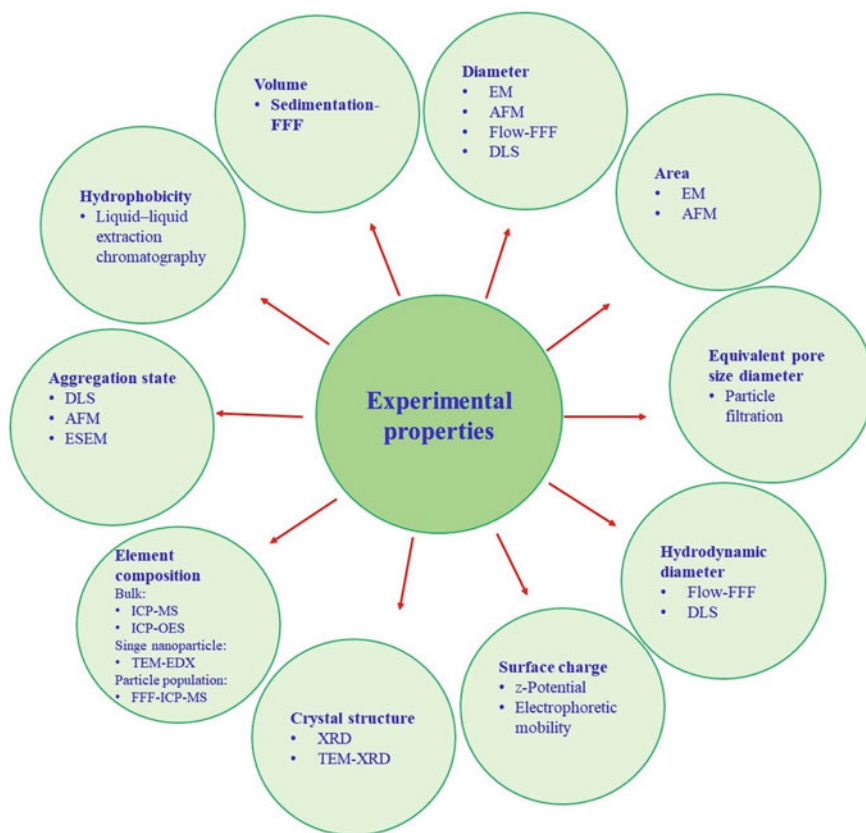


Fig. 2.3 Experimental characteristics as descriptors in nano-QSAR research [110]

the SMILES symbol. However, one should also be aware of their significant differences [121]. The SMILES can be produced by popular software such as ChemSketch, Biovia, and Chem Draw [122].

The prediction of activity/property of nanomaterials can be predicted by SMILES [123–125]. Quasi-SMILES is an alternative of SMILES-based optimal descriptors to build up predictive models for nanomaterials and other materials by consideration of the experimental conditions. Quasi-SMILES may be eclectic condition [126, 127] or combination of SMILES and eclectic conditions [128, 129]. The continuous eclectic conditions can be normalized by the following equation for assigning codes:

$$\text{Norm}(P_i) = \frac{\min(P_i) + P_i}{\min(P_i) + \max(P_i)} \quad (2.18)$$

P_i is its value of physicochemical parameter P , $\min(P_i)$ is minimum value of P and $\max(P_i)$ indicates maximum value of P .

Table 2.2 Distinction of standardized physiochemical features into classes 1–9 according to its value

Norm value	Class
$\text{Norm}(P) > 0.9$	9
$0.8 < \text{Norm}(P) < 0.9$	8
$0.7 < \text{Norm}(P) < 0.8$	7
$0.7 < \text{Norm}(P) < 0.6$	6
$0.6 < \text{Norm}(P) < 0.5$	5
$0.5 < \text{Norm}(P) < 0.4$	4
$0.4 < \text{Norm}(P) < 0.3$	3
$0.3 < \text{Norm}(P) < 0.2$	2
$0.2 < \text{Norm}(P) < 0.1$	1
$\text{Norm}(P) < 0.1$	0

According to Table 2.2, the number of unique values in each parameter was less than 10; therefore, the quasi-SMILES descriptors representations could be coded by assigning a number between zero and nine in a single character.

2.3.1 Quasi-SMILES Examples in Peer-Reviewed Papers

Table 2.3 shows an example of the construction codes for the quasi-SMILES. Based on the data shown in Table 2.3, the quasi-SMILES can be generated, which can be used to build a model according to the optimal descriptors. Table 2.4 indicates some examples for quasi-SMILES generated by codes shown in Table 2.3.

The new reported QSPR analysis of MOFs by Ahmadi et al. is application of quasi-SMILES parameters including Brunauer, Emmett, and Teller (BET) specific surface area and pore volume, pressure, and temperature for prediction of CO₂ adsorption of MOFs [128]. Tables 2.5 and 2.6 show the eclectic data range and quasi-SMILES codes for them, respectively.

In the code-2019 of CORAL software for quasi-SMILES groups of symbols %10–%99 (reserved for representation of complex systems of rings for usual SMILES) were applied as codes for the quasi-SMILES (Table 2.6). The disadvantage of this version of quasi-SMILES is the difficulty of interpretation of results by a user.

Further development of the CORAL software (CORAL-2020) allows the display of experimental conditions through groups of symbols enclosed in parentheses. Table 2.7 shows the comparison codes in the last version (CORAL-2020) and old version of CORAL for creating quasi-SMILES in recently proposed models for the mutagenic potential. One can see codes-2020 are quite transparent and consequently are more convenient for a user. As is clearly evident, CORAL-2020 codes are quite transparent and thus more user-friendly.

Table 2.3 Codes used for the cell line, method, time exposition, concentration, size of nanoparticles, and type of metal oxide to convert various information of experimental data into quasi-SMILES [126]

Feature	Value or type	Code	Feature	Value or type	Code
Cell line	MCF-7	H	Normalized nanoparticles size	$0.2 < \text{Norm}(\text{size}) \leq 0.3$	P
	HT-1080	I		$0.3 < \text{Norm}(\text{size}) \leq 0.4$	Q
	HepG-2	J		$0.4 < \text{Norm}(\text{size}) \leq 0.5$	R
	HT-29	K		$0.5 < \text{Norm}(\text{size}) \leq 0.6$	S
	PC-12	L		$0.9 < \text{Norm}(\text{size}) \leq 1.0$	T
Method	MTT	M	Metal oxide type	SnO ₂	1
	NRU	N		MnO ₂	2
Time exposition	24	X		ZnO	3
	48	Y		Bi ₂ O ₃	4
	72	Z		NiO	5
Concentration (μg mL ⁻¹)	5	A		CeO ₂	6
	10	B		SiO ₂	7
	25	C		TiO ₂	8
	50	D			
	100	E			
	200	F			

Toropov et al. reported the model of toxicity examined based on four eclectic data including three possible forms of silver nanoparticles (bare, coat, cons), organisms (*Daphnia magna* or Zebrafish), size (nm), and zeta-potential (mV) [131], where “bare” characterizes nanoparticles without any coating, coat (coating) demonstrates nanoparticles with a shell, and “cons” defines nanoparticles including coating material descriptors (Table 2.8).

2.4 Software for Generation of Molecular Descriptors

Over the last two decades, the growing interest in property/activity prediction has led to the release of many software products to the market and open-source domains for scientists working in the field of QSPR/QSAR modeling. Table 2.9 shows some popular software for calculating molecular descriptors. In addition, some of them are complex packages that also include modules for QSPR/QSAR modeling, statistical analysis, and data visualization.

Table 2.4 Some examples for quasi-SMILES produced by codes indicated in Table 2.3

Cell line	Method	Time exposition (h)	Concentration ($\mu\text{g mL}^{-1}$)	Normalized NPs size	Metal oxide type	Quasi-SMILES	Cell viability (%)
MCF-7	MTT	24	10	$0.2 < \text{Norm}(\text{size}) \leq 0.3$	SnO ₂	HMXBP1	94.5
MCF-7	MTT	24	10	$0.2 < \text{Norm}(\text{size}) \leq 0.3$	MnO ₂	HMXBP2	91.0
HT-1080	NRU	24	25	$0.2 < \text{Norm}(\text{size}) \leq 0.3$	MnO ₂	INXCP2	79.0
MCF-7	MTT	48	50	$0.2 < \text{Norm}(\text{size}) \leq 0.3$	ZnO	HMYDP3	5.0
HepG-2	NRU	72	5	$0.2 < \text{Norm}(\text{size}) \leq 0.3$	SiO ₂	JNZAP7	95.7
PC-12	MTT	48	50	$0.4 < \text{Norm}(\text{size}) \leq 0.5$	TiO ₂	LMYDR8	52.0
HT-1080	MTT	24	25	$0.5 < \text{Norm}(\text{size}) \leq 0.6$	NiO	IMXCS5	77.0
HT-29	MTT	24	100	$0.3 < \text{Norm}(\text{size}) \leq 0.4$	CeO ₂	KMXEQ6	88.5
MCF-7	MTT	24	100	$0.3 < \text{Norm}(\text{size}) \leq 0.4$	NiO	HMXEQ5	51.7

Table 2.5 Lower and high levels of CO₂ capture capacity, BET, pore volume, pressure (bar), and temperature (K) [128]

	CO ₂ capture capacity (mol/kg)	BET	Pore volume (cm ³ /g)	Pressure (bar)	Temperature (K)
Low level	0.1	0	0.035	0.01	195
High level	54.5	6240	7.5	55	318

Table 2.6 Defined quasi-SMILES codes for eclectic conditions (BET-normalized, normalized pore volume normalized, pressure-normalized, and temperature-normalized) of CO₂ capture capacity of MOFs [128]

Normalized range	BET	Code-2019 for pore volume	Code-2019 for pressure	Code-2019 for temperature
0 < BET – normalized ≤ 0.1	%10	%20	%30	%40
0.1 < BET – normalized ≤ 0.2	%11	%21	%31	%41
0.2 < BET – normalized ≤ 0.3	%12	%22	%32	%42
0.3 < BET – normalized ≤ 0.4	%13	%23	%33	%43
0.4 < BET – normalized ≤ 0.5	%14	%24	%34	%44
0.5 < BET – normalized ≤ 0.6	%15	%25	%35	%45
0.6 < BET – normalized ≤ 0.7	%16	%26	%36	%46
0.7 < BET – normalized ≤ 0.8	%17	%27	%37	%47
0.8 < BET – normalized ≤ 0.9	%18	%28	%38	%48
0.9 < BET – normalized ≤ 1	%19	%29	%39	%49

2.5 Conclusion and Future Direction

Molecular descriptors are a critical component of the methodological toolbox used to study quantitative structure–property/activity relationship (QSPR/QSAR) modeling and are widely used to describe the structures of chemical compounds for design of new compounds. The predictive and reliable QSPR/QSAR models depend on accurate descriptors, as accurate predictions can save the time and cost needed to design new compounds with the desired property/activity.

In this chapter, the main classes of theoretical molecular descriptors including 0D, 1D, 2D, 3D, and 4D descriptors are described. The most significant progress

Table 2.7 Definition of eclectic condition for the definition of quasi-SMILES [130]

	Condition	Code-2019	Code-2020
Coating	TA100	%10	[TA100]
	TA98	%11	[TA98]
	20-nm citrate	%12	[20cit]
	20-nm PVP	%13	[20PVP]
	50-nm citrate	%14	[50cit]
	50-nm PVP	%15	[50PVP]
	100-nm citrate	%16	[100cit]
Doses ($\mu\text{g}/\text{plate}$)	100-nm PVP	%17	[100PVP]
	0.0	%18	[d0.0]
	6.3	%19	[d6.3]
	12.5	%20	[d12.5]
	25	%21	[d25]
	50	%22	[d50]
	100	%23	[d100]

Table 2.8 Indicates some quasi-SMILES used to generate nano-QSAR model for pLC₅₀ [131]

Status of nanoparticles	Organisms	Size (nm)	Zeta-potential (mV)	Quasi-SMILES
nanoparticles without any coating	Daphnia magna	17.150–21.700	– 8.480 to – 5.050	[Bare][Daph][s%14][z%25]
NPs without any coating	Daphnia magna	12.600–17.150	– 25.630 to – 22.200	[Bare][Daph][s%13][z%20]
NPs with a shell	Daphnia magna	53.550–58.100	– 11.910 to – 8.480	[Daph][s%22][z%24]
NPs including coating material descriptors	Daphnia magna	21.700–26.250	– 11.910 to – 8.480	[Daph][s%15][z% 24]
NPs without any coating	Zebrafish	135.450–140.000	– 22.200 to – 18.770	[Bare][Fish][s%40][z%21]
NPs with a shell	Zebrafish	44.450–49.000	– 25.630 to – 22.200	[Fish][s%20][z%20]

over the last few years in chemometrics, cheminformatics, and bioinformatics has led to new strategies for finding new molecular descriptors. Here, some of the most common molecular descriptors and some new molecular descriptors especially for design and QSPR/QSAR modeling of nanocomposites have been highlighted.

In nano-QSPR/QSAR modeling, the data in many different publications are small and not ready enough for model building. In addition, nanomaterials exhibit high complexity and heterogeneity in their structures, which makes data collection and processing more challenging compared to traditional QSPR/QSAR. Quasi-SMILES

descriptors are one of the solutions to this challenge and have been introduced as new descriptors combining SMILES and eclectic conditions. These novel descriptors provide transparent interpretation equation models with correlation weights calculated by Monte Carlo optimization using CORAL software.

Finally, a list of the most commonly used software packages for calculating molecular descriptors is reviewed here.

Table 2.9 List of software packages for the calculation of molecular descriptors

Name	Organization/institution	Availability	Descriptors	Platform/license
RDKit	GitHub	https://github.com/rdkit	> 200	Windows/Linux/Mac (freeware)
PaDELPy	University of Massachusetts Lowell	https://github.com/ecrl/padelpy	> 2500	Windows/Linux/Mac (freeware)
ADAPT	Pennsylvania State University	http://research.chem.psu.edu/pcjgroup/adapt.html	> 260	Unix/Linux (freeware)
ADMET	Simulations Plus, Inc	http://www.simulations-plus.com/	297	Windows (commercial)
Predictor™ CODESSA	Semichem	http://www.semichem.com/codessa/default.php	> 600	Windows/Linux (commercial)
DRAGON	Talete SRL	http://www.talete.mi.it/products/dragon_description.htm	4885	Windows/Linux (commercial)
EPISUITE™	EPA	http://www.epa.gov/opptintr/exposure/pubs/episuite.htm	20	Windows (freeware)
MOE	Chemical Computing Group	http://www.chemcomp.com/software-moe2009.htm	> 300	Windows/Linux/SGI/MAC/Sun (freeware)

(continued)

Table 2.9 (continued)

Name	Organization/institution	Availability	Descriptors	Platform/license
Molconn-Z™	EduSoft	http://www.edusoft-lc.com/molconn/	327	Windows/Unix/MAC (commercial)
MOLD	NCTR/FDA	http://www.fda.gov/ScienceResearch/BioinformaticsTools/Mold2/default.htm	777	Windows (freeware)
MOLGEN	University of Bayreuth	http://www.molgen.de/?src¼documents/molgenqspr.html	707	Windows (commercial)
PowerMV	NISS	https://www.niss.org/research/software/powermv	> 1000	Windows (freeware)
Sarchitect™	Strand Life Sciences	http://www.strandls.com/sarchitect/index.html	1084	Windows/Linux (commercial)
SciQSAR™	SciMatics	http://www.scimatics.com/jsp/qsar/QSARIS.jsp	> 600	Windows (commercial)
Alvadesc	Alvascience	https://www.alvascience.com/alvadesc/	> 6000	Windows/Linux/MAC (commercial)
CORAL	Istituto di Ricerche Farmacologiche Mario Negri	http://www.insilico.eu/coral/SOFTWARECORAL.html	> 1000	Windows (freeware)

References

1. Rocke AJ (1981) *Br J Hist Sci* 14:27–57. <https://doi.org/10.1017/S0007087400018276>
2. Kekulé A (858) *Liebigsder Chemie JA* 106:129–159. <https://doi.org/10.1002/jlac.18581060202>
3. Brown AC, Fraser TR (1868) *Eearth Environ Sci Trans Roy Soc* 25:151–203. <https://doi.org/10.1017/S0080456800028155>
4. Richardson B (1869) *Med Times Gazzette* (ii), pp 703–706
5. Körner W (1874) *Gazz Chim It* 4:242
6. Richet M (1893) *Compt Rend Soc Biol (Paris)* 45:775–776
7. Hammett LP (1937) *J Am Chem Soc* 59:96–103. <https://doi.org/10.1021/ja01280a022>
8. Ghasemi J, Ahmadi S (2007) *Ann Chim* 97:69–83. DOI:<https://doi.org/10.1002/adic.200690087>
9. Ahmadi S, Mardinia F, Azimi N, Qomi M, Balali E (2019) *J Mol Struct* 1181:305–311. <https://doi.org/10.1016/j.molstruc.2018.12.089>
10. Ahmadi S, Ghanbari H, Lotfi S, Azimi N (2021) *Mol Divers* 25:87–97. <https://doi.org/10.1007/s11030-019-10026-9>
11. Javidfar M, Ahmadi S (2020) *SAR QSAR Environ Res* 31:717–739. <https://doi.org/10.1080/1062936X.2020.1806922>
12. Ahmadi S (2012) *Macroheterocycles* 5:23–31. <https://doi.org/10.6060/mhc2012.110734a>
13. Ahmadi S, Babae E (2014) *J Incl Phenom Macro* 79:141–149. <https://doi.org/10.1007/s10847-013-0337-7>
14. Ahmadi S, Deligeorgiev T, Vasilev A, Kubista M (2012) *Russ J Phys Chem A* 86:1974–1981. <https://doi.org/10.1134/S0036024412130201>
15. Ghasemi JB, Ahmadi S, Brown S (2011) *Environ Chem Lett* 9:87–96. <https://doi.org/10.1007/s10311-009-0251-9>
16. Ahmadi S, Khazaei MR, Abdolmaleki A (2014) *Med Chem Res* 23:1148–1161. <https://doi.org/10.1007/s00044-013-0716-z>
17. Ahmadi S, Habibpour E (2017) *Anti-Cancer Agent Med Chem* 17:552–565. <https://doi.org/10.2174/1871520611009010001>
18. Ahmadi S (2012) *J Incl Phenom Macro* 74:57–66. <https://doi.org/10.1007/s10847-010-9881-6>
19. Ghasemi JB, Ahmadi S, Ayati M (2010) *Macroheterocycles* 3:234–242. <https://doi.org/10.6060/mhc2010.4.234>
20. Tropsha A, Wang S (2007) In: Bourne H, Horuk R, Kuhnke J, Michel H (eds) *GPCRs: from deorphanization to lead structure identification*. Ernst Schering Foundation symposium proceedings, vol 2006/2. Springer, Berlin, Heidelberg, pp 49–74. https://doi.org/10.1007/2789_2006_003
21. Ahmadi S, Ganji S (2016) *Curr Drug Discov Technol* 13:232–253. <https://doi.org/10.2174/1570163813666160725114241>
22. Lotfi S, Ahmadi S, Kumar P (2021) *J Mol Liq* 338:116465. <https://doi.org/10.1016/j.molliq.2021.116465>
23. Habibpour E, Ahmadi S (2017) *Curr Comput-Aid Drug Des* 13:143–159. <https://doi.org/10.2174/1573409913666170124100810>
24. Lotfi S, Ahmadi S, Kumar P (2021) *RSC Adv* 11:33849–33857. <https://doi.org/10.1039/D1R A06861J>
25. Willett P, Barnard JM, Downs GM (1998) *J Che Inf Comp Sci* 38:983–996. <https://doi.org/10.1021/ci9800211>
26. Suhachev D, Pivina T, Shlyapochnikov V, Petrov E, Palyulin V, Zefirov N (1993) *Dokl RAN* 328:50–57
27. Harary F (1971) *Graph theory*, 2nd printing. Addison-Wesley, Reading, MA
28. Roy K (2004) *Mol Divers* 8:321–323. <https://doi.org/10.1023/b:modi.0000047519.35591.b7>
29. Wiener H (1947) *J Am Chem Soc* 69:17–20. <https://doi.org/10.1021/ja01193a005>
30. Randić M (1993) *Chem Phys Lett* 211:478–483. [https://doi.org/10.1016/0009-2614\(93\)87094-J](https://doi.org/10.1016/0009-2614(93)87094-J)

31. Nikolić S, Trinajstić N, Randić M (2001) *Chem Phys Lett* 333:319–321. [https://doi.org/10.1016/S0009-2614\(00\)01367-1](https://doi.org/10.1016/S0009-2614(00)01367-1)
32. Li X-h, Li Z-g, Hu M-l (2003) *J Mol Graph Model* 22:161–172. [https://doi.org/10.1016/S1093-3263\(03\)00157-8](https://doi.org/10.1016/S1093-3263(03)00157-8)
33. Randić M (1975) *J Am Chem Soc* 97:6609–6615. <https://doi.org/10.1021/ja00856a001>
34. Kier LB, Hall LH (1986) In: *Molecular connectivity in structure-activity analysis. Research studies*. Wiley, Letchworth, Hertfordshire, England, New York, p 262. <https://doi.org/10.1002/jps.2600760325>
35. Randić M (1991) *J Mat Chem* 7:155–168. <https://doi.org/10.1007/BF01200821>
36. Randić M (2001) *J Mol Graph Model* 20:19–35. [https://doi.org/10.1016/S1093-3263\(01\)00098-5](https://doi.org/10.1016/S1093-3263(01)00098-5)
37. Kier L (1986) *Acta Pharm Jugosl* 36:171–188. <https://doi.org/10.1002/med.2610070404>
38. Balaban AT (1982) *Chem Phys Lett* 89:399–404. [https://doi.org/10.1016/0009-2614\(82\)80009-2](https://doi.org/10.1016/0009-2614(82)80009-2)
39. Gutman I, Das KC (2004) *MATCH Commun Math Comput Chem* 50:83–92. https://match.pmf.kg.ac.rs/electronic_versions/Match50/match50_83-92.pdf
40. Furtula B, Graovac A, Vukičević D (2010) *J Mat Chem* 48:370–380. <https://doi.org/10.1007/s10910-010-9677-3>
41. Hosoya H (1971) *B Chem Soc Jpn* 44:2332–2339. <https://doi.org/10.1246/bcsj.44.2332>
42. Randić M, Zupan J (2001) *J Chem Inf Comp Sci* 41:550–560. <https://doi.org/10.1021/ci00095o>
43. Moreau G, Broto P (1980) *Nouv J Chim* 4(6):359–360
44. Gutman I (1994) *Graph Theory Notes NY* 27(9):9–15
45. Melville JL, Hirst JD (2007) *J Chem Inf Model* 47:626–634. <https://doi.org/10.1021/ci6004178>
46. Spowage BM, Bruce CL, Hirst JD (2009) *J Cheminform* 1:1–13. <https://doi.org/10.1186/1758-2946-1-22>
47. Ghose AK, Crippen GM (1986) *J Comput Chem* 7:565–577. <https://doi.org/10.1002/jcc.540070419>
48. Arnott JA, Kumar R, Planey SL (2013) *J Appl Biopharm Pharmacokinet* 1:31–36. <http://creativecommons.org/licenses/by-nc/3.0/>
49. Winiwarter S, Ax F, Lennernäs H, Hallberg A, Pettersson C, Karlén A (2003) *J Mol Graph Model* 21:273–287. [https://doi.org/10.1016/S1093-3263\(02\)00163-8](https://doi.org/10.1016/S1093-3263(02)00163-8)
50. Cramer RD, Patterson DE, Bunce JD (1988) *J Am Chem Soc* 110:5959–5967. <https://doi.org/10.1021/ja00226a005>
51. Seel M, Turner DB, Willett P (1999) *Quant Struct-Act Relat* 18:245–252. [https://doi.org/10.1002/\(SICI\)1521-3838](https://doi.org/10.1002/(SICI)1521-3838)
52. Doweyko AM (1988) *J Med Chem* 31:1396–1406. <https://doi.org/10.1021/jm00402a025>
53. Kuz'min VE, Artemenko AG, Kovdienko NA, Tetko IV, Livingstone DJ (2000) *J Mol Model* 6:517–526. <https://doi.org/10.1007/s0089400060517>
54. Todeschini R, Consonni V (2008) *Handbook of molecular descriptors*. Wiley, p 667. <https://doi.org/10.1002/9783527613106>
55. Soltzberg LJ, Wilkins CL (1977) *J Am Chem Soc* 99:439–443. <https://doi.org/10.1021/ja00444a021>
56. Wagener M, Sadowski J, Gasteiger J (1995) *J Am Chem Soc* 117:7769–7775. <https://doi.org/10.1021/ja00134a023>
57. Vedani A, Dobler M (2000) In: Jucker E (ed) *Progress in drug research*, vol 55. Birkhäuser, Basel, pp 105–135. https://doi.org/10.1007/978-3-0348-8385-6_4
58. Easton RE, Giesen DJ, Welch A, Cramer CJ, Truhlar DG (1996) *Theor Chim Acta* 93:281–301. <https://doi.org/10.1007/BF01127507>
59. Kostal J, Voutchkova-Kostal A, Anastas PT, Zimmerman JB (2015) *Proc Natl Acad Sci* 112:6289–6294. <https://doi.org/10.1073/pnas.1314991111>
60. Lynch BJ, Truhlar DG (2004) *Theor Chem Acc* 111:335–344. <https://doi.org/10.1007/s00214-003-0518-3>

61. Azimi A, Ahmadi S, Kumar A, Qomi M, Almasirad A (2022) Polycycl Aromat Comp 1–21. <https://doi.org/10.1080/10406638.2022.2067194>
62. Cocchi M, Menziani MC, De Benedetti PG, Cruciani G (1992) Chemometr Intell Lab 14:209–224. [https://doi.org/10.1016/0169-7439\(92\)80105-D](https://doi.org/10.1016/0169-7439(92)80105-D)
63. Franke R (1984) Pharm Libr, vol 7. Elsevier, Amsterdam, p 412
64. Bodor N, Gabanyi Z, Wong CK (1989) J Am Chem Soc 111:3783–3786. <https://doi.org/10.1021/ja00193a003>
65. Buydens L, Massart DL, Geerlings P (1983) Anal Chem 55:738–744. <https://doi.org/10.1021/ac00255a034>
66. Klopman G, Iroff LD (1981) J Comput Chem 2:157–160. <https://doi.org/10.1002/jcc.540020204>
67. Zhou Z, Parr RG (1990) J Am Chem Soc 112:5720–5724. <https://doi.org/10.1021/ja00171a007>
68. Ośmiałowski K, Halkiewicz J, Radecki A, Kaliszan R (1985) J Chromatogr A 346:53–60. [https://doi.org/10.1016/S0021-9673\(00\)90493-X](https://doi.org/10.1016/S0021-9673(00)90493-X)
69. Fukui K (1970) In: Orientation and Stereoselection. Fortschritte der Chemischen Forschung, vol 15/1. Springer, Berlin, Heidelberg, pp 1–85. <https://doi.org/10.1007/BFb0051113>
70. Sklenar H, Jäger J (1979) Int J Quantum Chem 16:467–484. <https://doi.org/10.1002/qua.560160306>
71. Tuppurainen K, Lötjönen S, Laatikainen R, Vartiainen T, Maran U, Strandberg M, Tamm T (1991) Mutat Res Fundam Mol Mech 247:97–102. [https://doi.org/10.1016/0027-5107\(91\)90037-O](https://doi.org/10.1016/0027-5107(91)90037-O)
72. Becker H (1978) J Prakt Chem 320:879–880. <https://doi.org/10.1002/prac.19783200525>
73. Lewis D, Ioannides C, Parke D (1994) Xenobiotica 24:401–408. <https://doi.org/10.3109/00498259409043243>
74. Pearson RG (1989) J Org Chem 54:1423–1430. <https://doi.org/10.1021/jo00267a034>
75. Prabhakar YS (1991) Drug Des Deliv 7:227–239
76. Kurtz HA, Stewart JJ, Dieter KM (1990) J Comput Chem 11:82–87. <https://doi.org/10.1002/jcc.540110110>
77. Cammarata A (1967) J Med Chem 10:525–527. <https://doi.org/10.1021/jm00316a004>
78. Leo A, Hansch C, Church C (1969) J Med Chem 12:766–771. <https://doi.org/10.1021/jm00305a010>
79. Hansch C, Coats E (1970) J Pharm Sci 59:731–743. <https://doi.org/10.1002/jps.2600590602>
80. Lewis DF (1987) J Comput Chem 8:1084–1089. <https://doi.org/10.1002/jcc.540080803>
81. Cartier A, Rivail J-L (1987) Chemometr Intell Lab 1:335–347. [https://doi.org/10.1016/0169-7439\(87\)80039-4](https://doi.org/10.1016/0169-7439(87)80039-4)
82. Gaudio AC, Korolkovas A, Takahata Y (1994) J Pharm Sci 83:1110–1115. <https://doi.org/10.1002/jps.2600830809>
83. Grunenberg J, Herges R (1995) J Chem Inf Comp Sci 35:905–911. <https://doi.org/10.1021/ci00027a018>
84. Kikuchi O (1987) Quant Struct-Act Relat 6:179–184. <https://doi.org/10.1002/qsar.19870060406>
85. Grüber C, Buss V (1989) Chemosphere 19:1595–1609. [https://doi.org/10.1016/0045-6535\(89\)90503-1](https://doi.org/10.1016/0045-6535(89)90503-1)
86. Saura-Calixto F, Garcia-Raso A, Raso M (1984) J Chromatogr Sci 22:22–26. <https://doi.org/10.1093/chromsci/22.1.22>
87. Shusterman A (1991) ChemTech 21(10):624–627
88. Brusick DJ, Vogel EW, Nivard MJ, Klopman G, Rosenkranz HS, Enslein K, Gombar VK, Blake BW, Debnath AK, Shusterman AJ, de Compadre RL (1994) Mutat Res 305:321–323
89. Trapani G, Carotti A, Franco M, Latrofa A, Genchi G, Liso G (1993) Eur J Med Chem 28:13–21. [https://doi.org/10.1016/0223-5234\(93\)90074-O](https://doi.org/10.1016/0223-5234(93)90074-O)
90. Ebert C, Linda P, Alunni S, Clementi S, Cruciani G, Santini S (1990) Gazz Chim Ital 120:29
91. Skagerberg B, Bonelli D, Clementi S, Cruciani G, Ebert C (1989) Quant Struct-Act Relat 8:32–38. <https://doi.org/10.1002/qsar.19890080105>

92. Flynn GL (1980) *J Pharm Sci* 69:1109–1109. <https://doi.org/10.1002/jps.2600690938>
93. Taft RW, Topsom R (1987) *Prog Phys Org Chem* 16:1–83
94. Ewing DF (1979) *Org Magn Reson* 12:499–524. <https://doi.org/10.1002/mrc.1270120902>
95. Waterbeemd H Van de, Testa B (1987) In: *Advances in drug research*, vol 16. Academic, London, pp 87–227
96. Yang GZ, Lien EJ, Guo ZR (1986) *Quant Struct-Act Relat* 5:12–18. <https://doi.org/10.1002/qsar.19860050104>
97. Kim KH, Martin YC (1991). In: Silipo C, Vittoria A (eds) *QSAR: rational approaches to the design of bioactive compounds*. Elsevier, Amsterdam, pp 151–154
98. Sjöström M, Wold S (1985) *J Mol Evol* 22:272–277. <https://doi.org/10.1007/BF02099756>
99. Sjöström M, Wold S, Wieslander A, Rilfors L (1987) *EMBO J* 6:823–831. <https://doi.org/10.1002/j.1460-2075.1987.tb04825.x>
100. Carlson R, Lundstedt T, Albano C (1985) *Acta Chem Scand B* 39:79–91. <https://doi.org/10.3891/acta.chem.scand.39b-0079>
101. Carlson R, Lundstedt T, Nordahl Å, Prochazka M (1986) *Acta Chem Scand B* 40:522–533. <https://doi.org/10.3891/acta.chem.scand.40b-0522>
102. Lundstedt T, Carlson R, Shabana R (1987) *Acta Chem Scand B* 41:157–163. <https://doi.org/10.3891/acta.chem.scand.41b-0157>
103. Carlson R, Prochazka M, Lundstedt T (1988) *Acta Chem Scand B Org Chem Biochem* 42:145–156. <https://doi.org/10.3891/acta.chem.scand.42b-0145>
104. Hellberg S, Sjoström M, Skagerberg B, Wikström C, Wold S (1987) *Acta Pharm Jugosl* 37:53–65. <https://doi.org/10.1021/jm00390a003>
105. Hansch C, Leo A, Unger SH, Kim KH, Nikaitani D, Lien EJ (1973) *J Med Chem* 16:1207–1216. <https://doi.org/10.1021/jm00269a003>
106. Verloop A, Hoogenstraaten W, Tipker J (1976) In: Ariens EJ (ed) *Drug design*. Academic, New York
107. Rybińska-Fryca A, Mikolajczyk A, Puzyn T (2020) *Nanoscale* 12:20669–20676. <https://doi.org/10.1039/D0NR05220E>
108. Richarz A-N, Avramopoulos A, Benfenati E, Gajewicz A, Golbamaki Bakhtyari N, Leonis G, Marchese Robinson RL, Papadopoulos MG, Cronin MTD, Puzyn T (2017) In: Tran L, Bañares M, Rallo R (eds) *Modelling the toxicity of nanoparticles*. *Advances in experimental medicine and biology*, vol 947. Springer, Cham, pp 303–324. https://doi.org/10.1007/978-3-319-47754-1_10
109. Puzyn T, Gajewicz A, Leszczynska D, Leszczynski J (2010) In: Puzyn T, Leszczynski J, Cronin M (eds) *Recent advances in QSAR studies*. *Challenges and advances in computational chemistry and physics*, vol 8. Springer, Dordrecht, pp 383–409. https://doi.org/10.1007/978-1-4020-9783-6_14
110. Hassellöv M, Readman JW, Ranville JF, Tiede K (2008) *Ecotoxicology* 17:344–361. <https://doi.org/10.1007/s10646-008-0225-x>
111. Ahmadi S, Lotfi S, Kumar P (2020) *SAR QSAR Environ Res* 31:935–950. <https://doi.org/10.1080/1062936X.2020.1842495>
112. Ghiasi T, Ahmadi S, Ahmadi E, Talei Babil Olyai M, Khodadadi Z (2021) *SAR QSAR Environ Res* 32:495–520. <https://doi.org/10.1080/1062936X.2021.1925344>
113. Ahmadi S, Lotfi S, Afshari S, Kumar P, Ghasemi E (2021) *SAR QSAR Environ Res* 32:1013–1031. <https://doi.org/10.1080/1062936X.2021.2003429>
114. Ahmadi S, Lotfi S, Kumar P (2022) *Toxicol Mech Method* 32:302–312. <https://doi.org/10.1080/15376516.2021.2000686>
115. Ahmadi S, Moradi Z, Kumar P, Almasirad A (2022) *J Recept Signal Transduct* 42:361–372. <https://doi.org/10.1080/10799893.2021.1957932>
116. Lotfi S, Ahmadi S, Kumar P (2022) *RSC Adv* 12:24988–24997. <https://doi.org/10.1039/D2R A03936B>
117. Weininger D (1988) *J Chem Inf Comp Sci* 28:31–36. <https://doi.org/10.1021/ci00057a005>
118. Weininger D, Weininger A, Weininger JL (1989) *J Chem Inf Comp Sci* 29:97–101. <https://doi.org/10.1021/ci00062a008>

119. Pinheiro GA, Mucelini J, Soares MD, Prati RC, Da Silva JL, Quiles MG (2020) *J Phys Chem A* 124:9854–9866. <https://doi.org/10.1021/acs.jpca.0c05969>
120. Lotfi S, Ahmadi S, Zohrabi P (2020) *Struct Chem* 31:2257–2270. <https://doi.org/10.1007/s11224-020-01568-y>
121. Toropov A, Toropova A, Martyanov S, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2011) *Chemometr Intell Lab* 109:94–100. <https://doi.org/10.1016/j.chemolab.2011.07.008>
122. Ahmadi S, Mehrabi M, Rezaei S, Mardafkan N (2019) *J Mol Struct* 1191:165–174. <https://doi.org/10.1016/j.molstruc.2019.04.103>
123. Ahmadi S, Akbari A (2018) *SAR QSAR Environ Res* 29:895–909. <https://doi.org/10.1080/1062936X.2018.1526821>
124. Heidari A, Fatemi MH (2017) *J Chin Chem Soc-Taip* 64:289–295. <https://doi.org/10.1002/jccs.201600761>
125. Kumar P, Kumar A (2020) *SAR QSAR Environ Res* 31:697–715. <https://doi.org/10.1080/1062936X.2020.1806105>
126. Ahmadi S (2020) *Chemosphere* 242:125192. <https://doi.org/10.1016/j.chemosphere.2019.125192>
127. Ahmadi S, Toropova AP, Toropov AA (2020) *Nanotoxicology* 14:1118–1126. <https://doi.org/10.1080/17435390.2020.1808252>
128. Ahmadi S, Ketabi S, Qomi M (2022) *New J Chem* 46:8827–8837. <https://doi.org/10.1039/D2NJ00596D>
129. Ahmadi S, Aghabeygi S, Farahmandjou M, Azimi N (2021) *Struct Chem* 32:1893–1905. <https://doi.org/10.1007/s11224-021-01748-4>
130. Toropov AA, Toropova AP (2019) *Sci Total Environ* 681:102–109. <https://doi.org/10.1016/j.scitotenv.2019.05.114>
131. Toropov AA, Kjeldsen F, Toropova AP (2022) *Chemosphere* 135086. <https://doi.org/10.1016/j.chemosphere.2022.135086>

Chapter 3

Application of SMILES to Cheminformatics and Generation of Optimum SMILES Descriptors Using CORAL Software



Andrey A. Toropov and Alla P. Toropova

Abstract This chapter uses a simplified molecular input-line entry system (SMILES) to solve diverse problems in science, technology, and medicine. SMILES can be useful to model quantitative structure–property/activity relationships (QSPRs/QSARs). The evolution of the applications of SMILES and the evolution of SMILES descriptors are discussed. The construction of so-called optimal descriptors based on SMILES using the CORAL software is described. These optimal descriptors are useful for training QSPR/QSAR models for a wide range of diverse properties.

Keywords QSPR/QSAR · SMILES · Quasi-SMILES · Variational autoencoders · SmilesDrawer · DeepSMILES

3.1 Introduction

Simplified molecular input-line entry system (SMILES) is a chemical notation system for chemical information processing. Weininger developed the SMILES system in 1988 [1–3]. It is based on principles of molecular graph theory and allows rigorous structure specification using minimal and natural grammar. SMILES is a line notation for representing molecular structure that is intuitive to chemists and also well suited for high-speed computer-based analysis. SMILES has an increasing number of database-related applications. Here we discuss the use of SMILES to train quantitative structure–property/activity relationship (QSPRs/QSARs) models.

There are several useful text-based line formalisms based on the molecular graph that has been applied to QSPR/QSAR analysis. Those include SMILES [4, 5], SMILES arbitrary target specification (SMARTS) [6, 7], International Chemical Identifier (InChI) [8–13]. SMILES is the most popular of these for the QSPR/QSAR community while the use of SMARTS [6, 7, 14] and InChI [15, 16] is much less

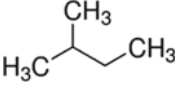
A. A. Toropov · A. P. Toropova (✉)

Laboratory of Environmental Chemistry and Toxicology, Department of Environmental Health Science, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156 Milano, Italy

e-mail: alla.toropova@marionegri.it

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. P. Toropova and A. A. Toropov (eds.), *QSPR/QSAR Analysis Using SMILES and Quasi-SMILES*, Challenges and Advances in Computational Chemistry and Physics 33, https://doi.org/10.1007/978-3-031-28401-4_3

Table 3.1 Examples of representation for 2-methyl butane

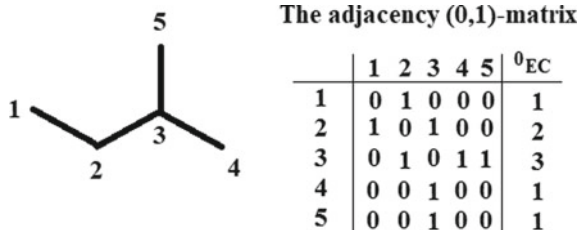
Structure	
SMILES	CC(C)CC
InChI	InChI=1/C5H12/c1-45(2)3/h5H,4H2,1-3H3
SMART	[#6]-[#6]-[#6](-[#6])-[#6]

common. The acronym SMARTS stands for SMILES arbitrary target specification. It is a language that allows specification of substructures for searching databases. Using SMARTS, flexible and efficient substructure-search specifications can be made in a way that is convenient for users. InChI, the International Chemical Identifier, also represents the molecular structure by sequences of special symbols.

The number of accessible internet molecular databases that use SMARTS and InChI representations is gradually increasing. This has accelerated the development models for physicochemical and/or biochemical endpoints based on SMARTS, SMILES, or InChI (i.e. directly from Internet databases). However, the number of models trained on SMARTS or InChI is still considerably smaller than those using SMILES. The main reason is that SMILES is a more natural and intuitive way to represent molecular structures for scientists. Table 3.1 contains examples of SMILES, InChI, and SMART for 2-methyl butane.

According to Einstein, “everything should be made as simple as possible, but not simpler” [17]. Despite SMILES being simpler than chemical graphs, historically, most of the descriptors used in practice are calculated using molecular graphs [18–29]. The molecular graph is a convenient representation of the molecular structure for the search for similarity and dissimilarity. This mathematical object has two categories of elements (i) vertexes (atoms) and (ii) edges (covalent bonds).

Wiener combined chemistry and mathematics in pioneering work on generating models for thermodynamic properties of paraffin compounds as a mathematical function of the molecular structure represented using the so-called the hydrogen-suppressed graph (HSG) [18–29]. The HSG can be expressed via the adjacency matrix, where 0 indicates the absence and 1 indicates the presence of a covalent bond between atoms. Figure 3.1 contains an example of the hydrogen-suppressed graph together with its equivalent adjacency matrix.

Fig. 3.1 Hydrogen-suppressed graph and the adjacency matrix for 2-methyl butane

The adjacency matrix is the basis for many topological indices [30–35]. Table 3.2 contains examples of topological indices calculated from the adjacency matrix. Researchers have generated many univariate and multivariate QSPR models for physicochemical endpoints using these as descriptors. Subsequently, Fujita et al. [30] established the first correlations for biochemical endpoints using these indices. The relationships between the molecular structure and biochemical effects can be more complex than between molecular structure and physicochemical properties. Metrics for the quality of those models were (i) the total number of compounds in the available set; (ii) correlation coefficient; (iii) root mean error or mean absolute error. The number of the topological indices and similar descriptors encoding physicochemical information increases exponentially with the size of molecules [36]. Unfortunately, in general, the increase in the quantity of indices and other descriptors derived from them is not accompanied by a rise in the quality of the corresponding models because of overfitting and other issues [37].

The uncertainty in model predictions even using superficial criteria was of concern to researchers, and new criteria for the statistical reliability of models were needed. Internal and external validation sets are commonly used to assess predictive power of models. The internal validation involves successively leaving or more molecules aside, calculating a model with the remainder, and using it to predict the properties of the molecules held aside [18]. When one molecule at a time is omitted (leave-one-out cross-validation, LOO), there is a low correlation between the external test set and LOO predictive [19]. The QSPR/QSAR model quality is heavily influenced by the type of molecular features used. Although the molecular graph was the mathematical representation of molecular features for building QSPR/QSAR models, SMILES [1–3, 20] can also represent molecular features.

The prediction of physicochemical and/or biochemical endpoints for a substance via computational procedures is an attractive alternative to the experimental measurement of the endpoints if this prediction is reliable. However, as machine learning

Table 3.2 Examples of topological indices (molecular descriptors) calculated from adjacency matrices [38]

Comment	Equations
Kier and Hall Zero-order connectivity index	${}^0X = \sum_k ({}^0EC_k)^{-1/2}$
Randic's connectivity index	${}^1X = \sum_{(k,j)\text{edge}} ({}^0EC_k \times {}^0EC_j)^{-1/2}$
Zagreb group index M1	$M1 = \sum_k ({}^0EC_k)^2$
Zagreb group index M2	$M2 = \sum_{(k,j)\text{edge}} ({}^0EC_k \times {}^0EC_j)$
Balaban index	$J = \frac{m}{\gamma+1} \sum_{(k,j)\text{edge}} ({}^0EC_k \times {}^0EC_j)$ where γ is the circuit rank of the graph, i.e. $\gamma = m - n + c$; m is the number of nodes; m is the number of edges; c is the number of cycles

methods used to generate QSAR models are data-driven, they are critically dependent on experimental data on similar substances with similar molecular structures being available.

Many studies have been dedicated to the similarity of substances. However, the best definition of molecular similarity is still not clear and is still controversial [38–40]. Comparative molecular similarity indices analysis (CoMSIA) and comparative molecular field analysis (CoMFA) provided important rational paradigms for molecular similarity [31]. The quality, quantity, and chemical diversity of experimental data used to train models is also very important if unbiased models with good predictive powers and lack of bias are to be achieved [33]. Molecular descriptors are also very important determinants of model quality and interpretability [34, 35]. Descriptors can be generated from molecular structure, physicochemical properties, biological properties, provenance properties or an other factors that may influence the property in question [36, 41]. Some descriptors are measured in experiments (e.g. octanol/water partition coefficient) but the most useful ones are generated mathematically experiment [37].

Thus, the SMILES representation of the molecular structure is visually and intuitively useful for perception and interpretation by users. In contrast, InChI or SMARTS text strings are less intuitive for people, despite (or perhaps because of) the higher levels of information available in an InChI. Modelling of complex phenomena almost always involves some simplification. Simplicity is a necessary and often useful abstraction in science, promoting clarity and interpretability at the expense of rigour. As best stated by Box: “All models are wrong, but some are useful” [42, 43]. The domain of applicability of models is also a very important and sometimes neglected property of QSAR models (indeed any ML models). This is the region of descriptor and property space spanned by the molecules in the training set. A large number of articles devoted to the applicability domain were written under the auspices of the Organisation for Economic Co-operation and Development (OECD) [44–46]. Clearly, models with ideal applicability domains that can generalize to any molecules in the whole of chemistry space are impossible. However, with appropriate choice of descriptors that encode all relevant molecular properties relevant to the property being modelled quite wide extrapolations of training chemical spaces are possible.

Molecular descriptors can be calculated in many ways, from high-level quantum chemical calculations or mathematical analyses of molecules, though descriptors derived from the chemical graph to those that are molecular fragment-, fingerprint-, or signature-based. Software packages that are easy to learn and produce useful results are clearly more popular with researchers.

SMILES has a growing cadre of users for solving diverse problems, especially since the advent of convolutional neural networks and related algorithms. Several other new descriptor generation and property modelling methods based on SMILES have been developed recently.

Variational autoencoders (VAEs) are a deep learning method designed to learn nonlinear latent representations which generalize to unseen data using SMILES [47–51]. A novel co-regularized variational autoencoders (Co-VAE) can predict drug-target binding affinity based on drug structures and target sequences. The Co-VAE model gives pairs VAEs for generating SMILES strings of therapeutically useful agents [47].

Deep neural networks effectively learn directly from low-level encoded data [52–57]. These models take the SMILES representation of the molecules as input to detect promising SMILES fragments from the latent descriptors they generate. In addition, the approach allows assessment of prediction uncertainty [58].

SmilesDrawer is a research tool capable of parsing and drawing SMILES-encoded molecular structures. It can display organic molecules in large numbers and fast succession. SmilesDrawer can draw structurally and stereochemically extremely complex structures [59, 60].

DeepSMILES is a recently proposed variant of SMILES, designed for rational analysis of extremely complex molecular structures. In addition, DeepSMILES propose useful simplifications to the SMILES syntax [61–63].

CurlySMILES: a chemical language to customize and annotate encodings of molecular and nanodevice structures [64] that is one more original modification of the traditional SMILES.

In 2015, an extension of traditional SMILES, quasi-SMILES, was proposed [64–69]. Quasi-SMILES extends standard SMILES but appends special codes for, for example, experimental conditions, and can also be used to generate models [65–70].

Quasi-SMILES can be used in the CORAL program discussed here, so we first describe this modification of traditional SMILES in more detail. The main reasons for developing quasi-SMILES are derived from the apparent analogy between the structure of peptides and ordinary molecules (in other words, for peptides we take amino acids as atoms) secondly, the strong influence of experimental conditions on various endpoints related to nanomaterials leads to attractively to apply special codes reflecting these conditions.

Initially, it was found that the correlation weighting of amino acids for peptides gives quite statistically significant results [65]. In addition, it was found that considering the experimental conditions significantly expands the possibilities for the development of nano-QSPR/QSAR [65–69].

The first attempts to construct quasi-SMILES used isolated symbols (1-hot descriptors) accounting for simple on/off effects (e.g. for the effect of lighting or heating). Later, special symbols were added to convey continuous properties (temperatures, solubility). Currently, users of the CORAL program can describe experimental conditions using identifiers (which include several characters for clarity) enclosed in square brackets [68–70]. Table 3.3 contains examples of quasi-SMILES.

For molecules, there is ambiguity in the sequence of symbols representing the molecule's structure, as valid SMILES can be defined starting from any point in the structure. This has been addressed by the concept of canonical SMILES, which are unique for every molecule. Developers of software generally enforce the use of generation of canonical SMILES for this reason.

Table 3.3 Examples of quasi-SMILES with the interpretation of components

Quasi-SMILES	Comment	References
X0 + A	The code 'X' means the presence of fullerene The code '0' means the presence of dark The code '+' means 'with Mix S9' The code 'A' means the dose 50 g/plate	[66]
[Bare][Daph][s%14][z%25]	[bare] = "nanoparticles without any coating" [daph] = " <i>Daphnia magna</i> " [s%14] = the range of size from 17.1 to 21.7 nm [z%25] = the range of zeta potential from - 8.48 to - 5.05 eV	[70]

3.1.1 The CORAL software description

Here we discuss a user-friendly program, CORAL (<http://www.insilico.eu/coral>). CORAL is an abbreviation of the words CORrelation And Logics. It aims to generate QSPR/QSAR models using input data lists of SMILES strings, together with corresponding experimental data on endpoints. In addition, this program can be applied to nano-QSPR/QSAR problems using lists of quasi-SMILES together with testing data on endpoints related to nanomaterials. In both mentioned cases, strings of symbols (SMILES or quasi-SMILES) are translated into the optimal descriptors.

A detailed description of the CORAL software follows that aims to provide a user with the necessary information on using the software without excess detail.

3.1.1.1 CORAL: Preparation of Input Files

CORAL website (<http://www.insilico.eu/coral>) contains several versions of the program. Previous versions may be convenient for users who have used them before (2016, 2017, 2019, and 2020) and may wish to apply them for similar new tasks. In addition, they provide the ability to verify and reproduce published models. However, only the program's latest version is described below since it contains all the features used in previous versions.

The standard name for the input file is "#TotalSet.txt". The file contains a list of the following strings:

ID...SMILES...Endpoint (here, three dots mean space).

ID

ID can be mean simple numbering 1, 2, 3, ..., N . In the case of research work dedicated to QSPR/QSAR analysis, the ID can be the chemical abstract service number (CAS) [71].

SMILES

Simplified molecular input-line entry system (SMILES) [1–3] is the widely used format for molecular structure representation. It is preferable to use canonical SMILES [72]. A popular software package to generate SMILES is ACD/ChemSketch [73, 74] although it is also easy to encode SMILES strings by hand if required.

Endpoint

There are no limitations for the endpoint for generating models using CORAL, but some rules should be considered. First, all compounds should express the endpoint in the same units. Second, ideally experimental data should be taken from one source. Third, the experimental conditions should be the same. For instance, solubility should relate to the same temperature, and toxicity should be associated with the same organisms, organs, and conditions.

3.1.1.2 CORAL: Selection of the Method

CORAL generates linear regression models expressed as:

$$\text{Endpoint} = \text{Intercept} + \text{Slope} * \text{Descriptor (SMILES)} \quad (3.1)$$

This model may appear extremely simple. However, the simplicity disappears after the task of the defining the calculation system for the optimal descriptor form SMILES is undertaken.

3.1.1.3 Defining the Optimal Descriptor

Figure 3.2 shows the interface of the CORAL program that defines how the optimum descriptors for QSAR/QSPR models are calculated. The complexity of this task arises from compromises between the information content of the selected molecular features extracted from SMILES and their representation across the training set. For example, the representation of molecules may be too detailed for the training set, resulting in an overfitted model that predicts the training set well but generalizes poorly. Structures outside the training set may also contain molecular features not in molecules in the training set. Conversely, if too few molecular features are used, then the model will be uninformative and predict both training and test sets poorly since a significant number of relevant molecular characteristics will be ignored when constructing the model. A suitable compromise can be reached using a simple logical trick or heuristic. The impact of molecular moieties from the simplest to the most complex are assessed by conducting appropriate computational experiments.

Below, the conception of the optimal descriptor is represented in more detail.

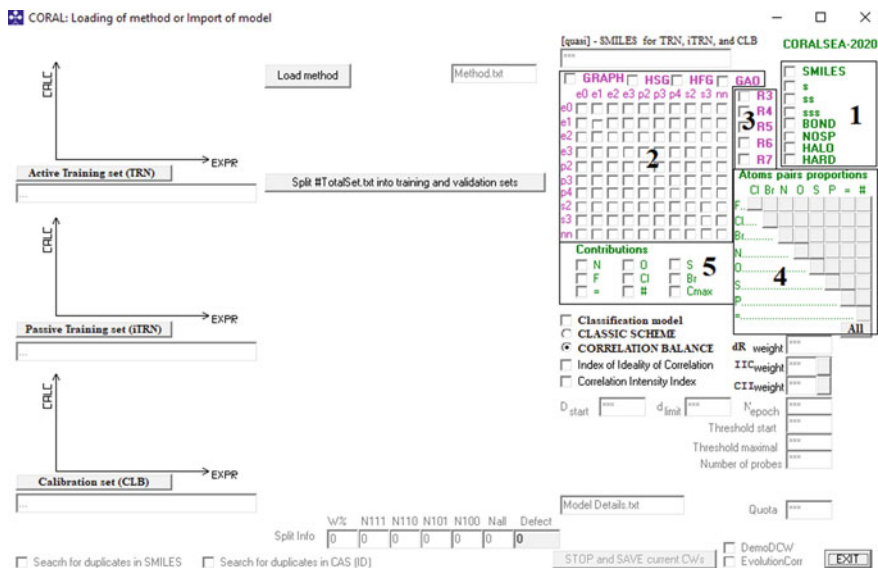


Fig. 3.2 Interface of the CORAL program to define the scheme of calculation of the optimal descriptor

1. Definition of SMILES components for the development of the optimal descriptor

Figure 3.3 shows part 1 of the interface in detail.

The user may ignore some contributions to the SMILES string, in which case the box relating to SMILES remains empty. If the user intends to use contributions coming directly from SMILES to construct an optimal descriptor, then the square referring to SMILES must be activated (Fig. 3.2).

In part 1, “s” denotes a “SMILES atom”, that is, a single character from the string SMILES (e.g. ‘C’, ‘N’, ‘O’, ‘=’, ‘#’, etc.) or a group of characters that cannot be considered in separately (e.g. ‘Cl’, ‘Br’, %11, [Zn], etc.). The “ss” denotes a pair of SMILES atoms following one after the other in the string SMILES (e.g. “CC”, “N1”,

Fig. 3.3 Choosing or not choosing to use SMILES attributes to calculate the optimal descriptor



SMILES attributes will not be used to calculate the optimal descriptor

SMILES attributes will be used to calculate the optimal descriptor

“C=”, “ON”, etc.). The “sss” denotes a triple of SMILES atoms following one after the other in the string SMILES (e.g. “CCC”, “O=C”, “=C2”, etc.).

It should be noted that it is unacceptable for pairs or triplets of SMILES atoms to be written in various sequences. For example, in one situation, “=C” is observed, and in the other, “C=”, since, in fact, both fragments represent the same situation in the molecule. The corresponding characters’ pairs and triples are fixed according to the ASCII codes [75] to avoid such inconsistencies.

s, ss, and sss are local SMILES attributes since they reflect the quality of local parts of SMILES strings.

BOND, HALO, NOSP, and HARD are global SMILES attributes since these reflect overall features of molecules extracted from SMILES. The BOND represents the presence or absence of different covalent bonds (double, triple, and stereo-chemical). The BOND is not sensitive to the numbers of these covalent bonds.

BOND

The BOND is built up as a configuration of twelve symbols. Table 3.4 contains examples of the twelve symbols in the BOND. Figure 3.4 contains graphical representations of the BOND attribute.

HALO

The HALO is a global SMILES attribute that reflects the presence or absence of fluorine, chlorine, bromine, and iodine atoms in a molecular structure. The HALO is a configuration of twelve symbols representing information on the above chemical elements. Table 3.5 contains simple examples of the HALO configurations. Figure 3.5 contains graphical representations for different statuses of the HALO attribute.

NOSP

The NOSP is a global SMILES attribute that reflects the presence or absence of nitrogen, oxygen, sulphur, and phosphorus atoms in a molecular structure. The NOSP is a configuration of twelve symbols representing information about the above chemical elements. Table 3.6 contains simple examples of the NOSP configurations. Figure 3.6 contains graphical representations of different statuses of the NOSP attribute.

HARD

In contrast to the above BOND, HALO, and NOSP, the global attribute HARD contains all the information in these attributes separately.

However, the information content of the HARD may be redundant if the training set is divided into many non-overlapping classes of molecular structures. Using BOND, HALO, and NOSP separately is a rational division of molecular structures into subclasses. Table 3.7 contains the general scheme of building up the twelve symbols code. Figure 3.7 shows the HARD configurations with the corresponding examples of the molecular structures.

Table 3.4 Definition of twelve symbols to encode the molecular physicochemical situation related to double, triple, and stereo-chemical bonds

1	2	3	4	5	6	7	8	9	10	11	12	Comment
B	O	N	D	0	0	0	0	0	0	0	0	Double, triple, and stereo-chemical (i.e. sensitive to three-dimensional geometry) bonds are absent in this molecule
B	O	N	D	1	0	0	0	0	0	0	0	The molecule contains double bonds (one or more) but does not contain triple or stereo-chemical bonds
B	O	N	D	0	1	0	0	0	0	0	0	The molecule contains triple bonds (one or more) but does not contain double bonds and stereo-chemical bonds
B	O	N	D	0	0	1	0	0	0	0	0	The molecule contains stereo-chemical bonds (one or more) but does not contain double and triple bonds

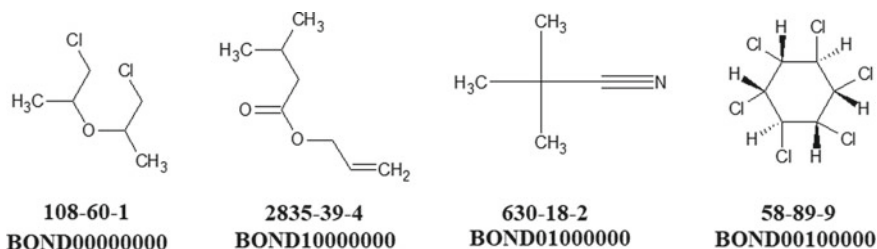


Fig. 3.4 Examples of different configurations for the BOND attribute

2. Definition of local and global graph invariants for the development of the optimal descriptor

SMILES and the molecular graph can be used to represent molecular structure data. Both approaches aim to represent molecular structure but have specific features so are not identical. This makes it tempting to compare the performance of these approaches for QSPR/QSAR analyses and to use both methods simultaneously in the hope of obtaining better results than those from the use of either of these approaches alone. Figure 3.8 contains the interface to select a group of different graph invariants.

The degree of the molecular graph vertex is the number of edges attached to this vertex. Figure 3.9 shows an example of a molecular graph. Having some (arbitrary) numbering, one can build up so-called adjacency (0, 1) matrix, where 1 means a covalent bond, and 0 indicates the absence of a bond for the corresponding pair of atoms in the graph. The adjacency matrix that gives possibility defines the sum of vertex degrees of neighbour atoms or defines the extended connectivity (Morgan extended connectivity [76]) (Fig. 3.10). A molecular graph built without considering hydrogen atoms is called a hydrogen-suppressed graph (HSG).

Note that the CORAL software allows use of the hydrogen-suppressed graph (HSG), the hydrogen-filled graph (HFG), and the graph of atomic orbitals (GAO) [75]. Figure 3.11 contains an example of GAO.

3. Accounting for the influence of molecular rings

The CORAL software interface allows the user with the opportunity to include the presence or absence of various rings. Figure 3.12 shows some examples of different versions of the use of the interface to take into account for the influence of molecular rings for building a model.

Special codes have been developed to account for the influence of molecular rings and other molecular features. Correlation weights are calculated that are used in calculating optimal descriptors. Figure 3.13 presents some examples of such codes that reflect the quality of the rings according to their size (3–7 membered rings), the presence (or absence) of heteroatoms, and the presence (or absence) of aromaticity.

Table 3.5 Definition of twelve symbols to encode the physicochemical situations related to halogens' presence (absence) in a molecule

1	2	3	4	5	6	7	8	9	10	11	12	Comment
				F	Cl	Br	I					
H	A	L	0	0	0	0	0	0	0	0	0	The molecule does not contain fluorine, chlorine, bromine, or iodine atoms
H	A	L	0	1	0	0	0	0	0	0	0	The molecule contains fluorine atoms (one or more) but does not contain chlorine, bromine, and iodine atoms
H	A	L	0	0	1	0	0	0	0	0	0	The molecule contains chlorine atoms (one or more) but does not contain fluorine, bromine, and iodine atoms
H	A	L	0	0	0	1	0	0	0	0	0	The molecule contains bromine atoms (one or more) but does not contain fluorine, chlorine, and iodine atoms
H	A	L	0	0	0	0	1	0	0	0	0	The molecule contains iodine atoms (one or more) but does not contain fluorine, chlorine, and bromine atoms

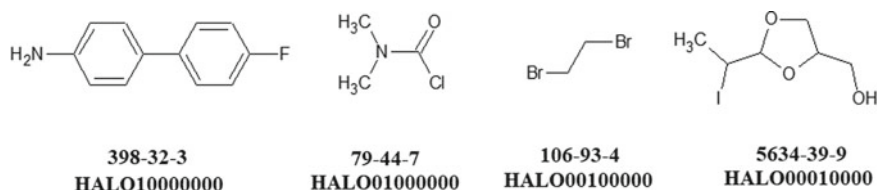


Fig. 3.5 Examples of different configurations for the HALO attribute

4. Atom pairs proportions (APP)

As discussed, molecular topology, i.e. the adjacency matrix of a molecular graph, provides detailed information on molecular structure. For modelling some, the ratio of various atoms can be important for building a model. In other words, biological activity may be affected by the ratio of the number of oxygen and nitrogen atoms or the ratio of the number of chlorine atoms and the double bonds [77].

To account for the influence of self-organizing vectors on the proportions of pairs of atoms when building a model, a fragment of the interface shown in Fig. 3.14 can be used.

5. Individual contributions of atoms

Descriptors or feature importance metrics are important for QSAR modelling because removing low relevance features and retaining only high relevance ones substantially improves model predictively and interpretability. One can test an atom (or several atoms from the list) for its ability to improve the statistical quality of the model and thus check whether the selected atom affects the predictive potential of the model or not. Figure 3.15 contains examples of applying the mentioned possibility.

6. Monte Carlo method algorithms

CORAL is a system of algorithms for building models and verifying them. There are several non-traditional methods for solving problems associated with modelling various endpoints. There is a long-held opposition between the ideas of determinism and randomness. It can be assumed that any model is a kind of random event, similar to the experimental observations of various physicochemical properties or biological activity.

CORAL uses random processes to build models that ensure the significance of reproducibility over the significance of accuracy for predictions of the model. These principles can be implemented in different ways.

- Classical QSPR/QSAR employs training and test sets. The information from the training set should be used to build a model, and the test set is being used to assess the ability of the model to generalize to unseen data.
- The balance of correlations method used active training, passive training sets, and some calibration set (an analogy of the test set). Figure 3.16 shows the difference between the classic scheme and the balance of correlations scheme.

Table 3.6 Definition of twelve symbols to encode the physicochemical situations related to the presence (absence) of nitrogen, oxygen, sulphur, and phosphorus

1	2	3	4	5	6	7	8	9	10	11	12	Comment
				N	O	S	P					
N	O	S	P	0	0	0	0	0	0	0	0	The molecule does not contain nitrogen, oxygen, sulphur, or phosphorus atoms
N	O	S	P	1	0	0	0	0	0	0	0	The molecule contains nitrogen atoms (one or more) but does not contain oxygen, sulphur, and phosphorus atoms
N	O	S	P	0	1	0	0	0	0	0	0	The molecule contains oxygen but does not contain nitrogen, sulphur, and phosphorus
N	O	S	P	0	0	1	0	0	0	0	0	The molecule contains sulphur, but does not contain nitrogen, oxygen, and phosphorus

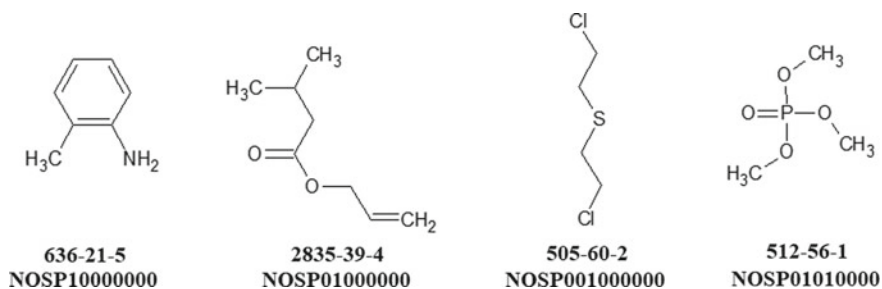


Fig. 3.6 Examples of different configurations for the NOSP attribute

The balance of correlations attempts to reduce the probability of a false model by employing the passive training set. A passive training set is a group of compounds similar to the active training set but with no overlap in these lists. In other words, the result of the traditional scheme can be expressed as “if the model is quite good for the training set, then one can expect that the model is good for external test set”. The result of the balance of correlations can be expressed as “the model is nice for active compounds which have been used to develop the model, and in addition, the model is not bad for compounds which are not used to develop the model”.

Computational experiments with the above two manners described in the literature confirm that the balance of correlations often gives better models than the traditional scheme ones [5, 78–82].

7. Monte Carlo optimization: its implementation and verification

The Monte Carlo method can be used to generate models using CORAL. The aim is to build an optimal descriptor capable of predicting endpoints through a regression relation of the form:

$$\text{EndPoint} = C_0 + C_1 \times \text{DCW}(T, N) \quad (3.2)$$

C_0 and C_1 are the regression coefficients. The T and N are special parameters governing the stochastic Monte Carlo optimization process. The T is the threshold for the definition of the active and blocked components of the optimal descriptor. If some component occurs in the training set (in the case of the balance of correlation, the active training set) more than T times, then it is active and is involved in building the model. If the indicated component occurs less than T times, it is blocked, and its correlation weight is equal to zero. Thus, a blocked component does not affect the model. The N is the number of iterations for Monte Carlo optimization. One iteration is a sequence of the modifications of all active components. The sequence of component modifications is random, and for each iteration, this sequence is determined anew.

The optimal descriptor is calculated as follows:

$$\text{DCW}(T, N) = \sum \text{CW}(\text{Component}_k) \quad (3.3)$$

Table 3.7 Definition of twelve symbols for encoding physical and chemical situations associated with the presence (absence) of various bonds (=, #, @), nitrogen, oxygen, sulphur, phosphorus, fluorine, chlorine, bromine, and iodine atoms in a molecule

1	2	3	4	5	6	7	8	9	10	11	12	Comment
=	#	@	N	O	S	P	F	Cl	Br	I		
\$	0	0	0	0	0	0	0	0	0	0	0	The molecule does not contain double, triple, stereo-chemical bonds, nitrogen, oxygen, sulphur, phosphorus, fluorine, chlorine, bromine, and iodine atoms
\$	1	0	0	1	1	0	0	0	0	1	0	The molecule contains double bonds (one or more), nitrogen, oxygen, and bromine atoms (one or several atoms each). Still, the molecule does not contain triple and stereo-chemical bonds, sulphur, phosphorus, fluorine, chlorine, and iodine atoms

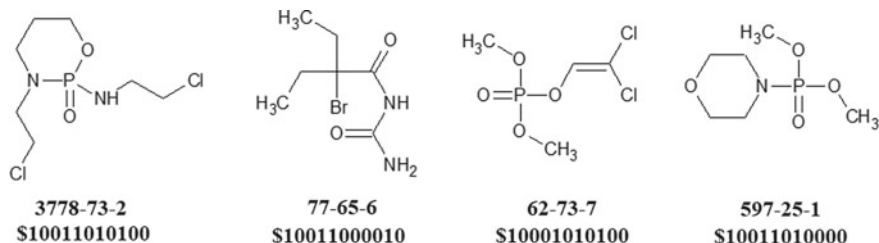


Fig. 3.7 Examples of different configurations for the HARD attribute

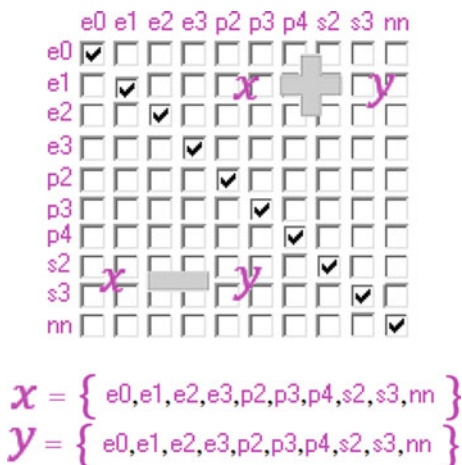
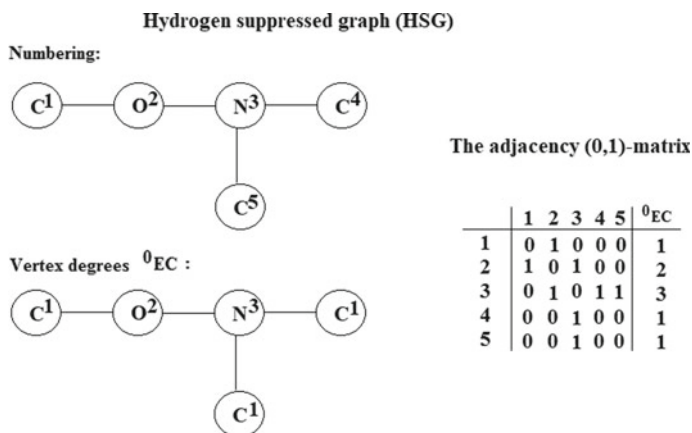
Fig. 3.8 Interface to select graph invariants: $e0$ = vertex degree; $e1$ – $e3$ = Morgan extended connectivity of first–third orders, respectively; $p2$ – $p4$ paths of lengths 2–4, respectively; $s2$, $s3$ = valence shells of second and third orders, respectively; nn = nearest neighbours codes

Fig. 3.9 Hydrogen-suppressed graph and the adjacency matrix

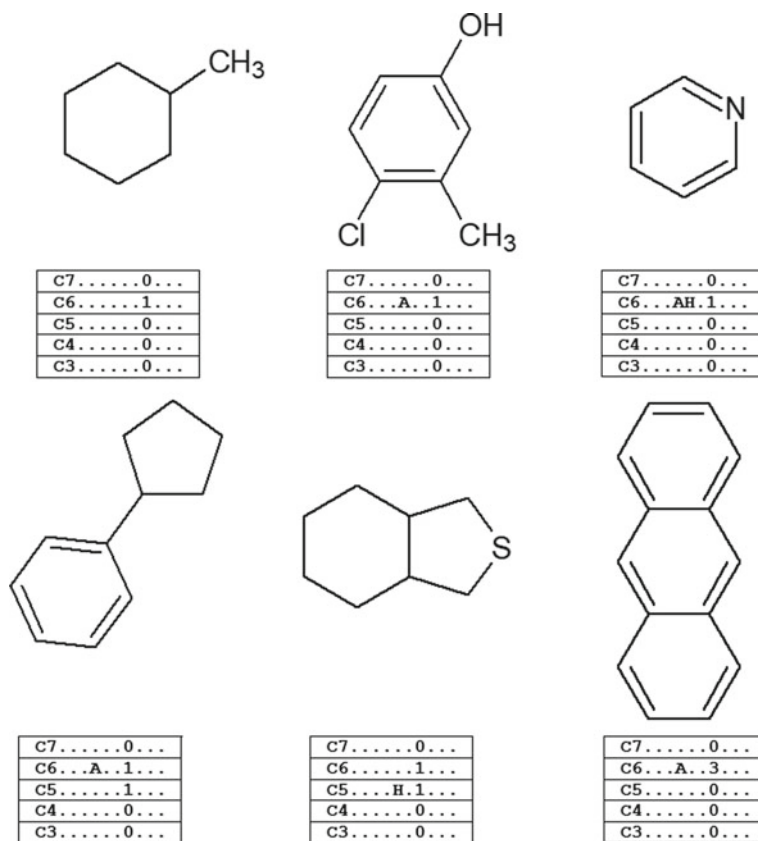


Fig. 3.13 Examples of codes applied to take into account the influence of molecular rings

The $CW(x)$ is the correlation weight of component x , obtained by the above Monte Carlo optimization. In addition to T and N , the optimization is controlled by special parametrization that defines the target function of the optimization:

$$\text{TargetFunction} = r_{AT} + r_{PT} - |r_{AT} - r_{PT}| \times \alpha + \text{IIC} \times \beta + \text{CII} \times \gamma \quad (3.4)$$

The r_{AT} and r_{PT} are correlation coefficients between the observed and predicted endpoint for the active and passive training sets, respectively. The IIC is the index of ideality of correlation [83–89]. The CII is the correlation intensity index [90–92]. The IIC is calculated with data on the calibration set as follows:

$$\text{IIC} = r \frac{\min(-\text{MAE}, +\text{MAE})}{\max(-\text{MAE}, +\text{MAE})} \quad (3.5)$$

Atoms pairs proportions	Atoms pairs proportions	Atoms pairs proportions																																																																																																																																																																																																																																																																														
<table border="1"> <thead> <tr> <th></th> <th>Cl</th> <th>Br</th> <th>N</th> <th>O</th> <th>S</th> <th>P</th> <th>=</th> <th>#</th> </tr> </thead> <tbody> <tr> <td>F.....</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>Cl.....</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>Br.....</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>N.....</td> <td></td> <td></td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>O.....</td> <td></td> <td></td> <td></td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>S.....</td> <td></td> <td></td> <td></td> <td></td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>P.....</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>=.....</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>0</td> </tr> <tr> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>All</td> </tr> </tbody> </table>		Cl	Br	N	O	S	P	=	#	F.....	0	0	0	0	0	0	0	0	Cl.....	0	0	0	0	0	0	0	0	Br.....	0	0	0	0	0	0	0	0	N.....			0	0	0	0	0	0	O.....				0	0	0	0	0	S.....					0	0	0	0	P.....						0	0	0	=.....								0									All	<table border="1"> <thead> <tr> <th></th> <th>Cl</th> <th>Br</th> <th>N</th> <th>O</th> <th>S</th> <th>P</th> <th>=</th> <th>#</th> </tr> </thead> <tbody> <tr> <td>F.....</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>Cl.....</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>Br.....</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>N.....</td> <td></td> <td></td> <td>2</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>O.....</td> <td></td> <td></td> <td></td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>S.....</td> <td></td> <td></td> <td></td> <td></td> <td>0</td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>P.....</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>0</td> <td>0</td> <td>0</td> </tr> <tr> <td>=.....</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>2</td> </tr> <tr> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>All</td> </tr> </tbody> </table>		Cl	Br	N	O	S	P	=	#	F.....	0	0	0	0	0	0	0	0	Cl.....	0	0	0	0	0	0	0	0	Br.....	0	0	0	0	0	0	0	0	N.....			2	0	0	0	0	0	O.....				0	0	0	0	0	S.....					0	0	0	0	P.....						0	0	0	=.....								2									All	<table border="1"> <thead> <tr> <th></th> <th>Cl</th> <th>Br</th> <th>N</th> <th>O</th> <th>S</th> <th>P</th> <th>=</th> <th>#</th> </tr> </thead> <tbody> <tr> <td>F.....</td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> </tr> <tr> <td>Cl.....</td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> </tr> <tr> <td>Br.....</td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> </tr> <tr> <td>N.....</td> <td></td> <td></td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> </tr> <tr> <td>O.....</td> <td></td> <td></td> <td></td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> </tr> <tr> <td>S.....</td> <td></td> <td></td> <td></td> <td></td> <td>2</td> <td>2</td> <td>2</td> <td>2</td> </tr> <tr> <td>P.....</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>2</td> <td>2</td> <td>2</td> </tr> <tr> <td>=.....</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>2</td> </tr> <tr> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>All</td> </tr> </tbody> </table>		Cl	Br	N	O	S	P	=	#	F.....	2	2	2	2	2	2	2	2	Cl.....	2	2	2	2	2	2	2	2	Br.....	2	2	2	2	2	2	2	2	N.....			2	2	2	2	2	2	O.....				2	2	2	2	2	S.....					2	2	2	2	P.....						2	2	2	=.....								2									All
	Cl	Br	N	O	S	P	=	#																																																																																																																																																																																																																																																																								
F.....	0	0	0	0	0	0	0	0																																																																																																																																																																																																																																																																								
Cl.....	0	0	0	0	0	0	0	0																																																																																																																																																																																																																																																																								
Br.....	0	0	0	0	0	0	0	0																																																																																																																																																																																																																																																																								
N.....			0	0	0	0	0	0																																																																																																																																																																																																																																																																								
O.....				0	0	0	0	0																																																																																																																																																																																																																																																																								
S.....					0	0	0	0																																																																																																																																																																																																																																																																								
P.....						0	0	0																																																																																																																																																																																																																																																																								
=.....								0																																																																																																																																																																																																																																																																								
								All																																																																																																																																																																																																																																																																								
	Cl	Br	N	O	S	P	=	#																																																																																																																																																																																																																																																																								
F.....	0	0	0	0	0	0	0	0																																																																																																																																																																																																																																																																								
Cl.....	0	0	0	0	0	0	0	0																																																																																																																																																																																																																																																																								
Br.....	0	0	0	0	0	0	0	0																																																																																																																																																																																																																																																																								
N.....			2	0	0	0	0	0																																																																																																																																																																																																																																																																								
O.....				0	0	0	0	0																																																																																																																																																																																																																																																																								
S.....					0	0	0	0																																																																																																																																																																																																																																																																								
P.....						0	0	0																																																																																																																																																																																																																																																																								
=.....								2																																																																																																																																																																																																																																																																								
								All																																																																																																																																																																																																																																																																								
	Cl	Br	N	O	S	P	=	#																																																																																																																																																																																																																																																																								
F.....	2	2	2	2	2	2	2	2																																																																																																																																																																																																																																																																								
Cl.....	2	2	2	2	2	2	2	2																																																																																																																																																																																																																																																																								
Br.....	2	2	2	2	2	2	2	2																																																																																																																																																																																																																																																																								
N.....			2	2	2	2	2	2																																																																																																																																																																																																																																																																								
O.....				2	2	2	2	2																																																																																																																																																																																																																																																																								
S.....					2	2	2	2																																																																																																																																																																																																																																																																								
P.....						2	2	2																																																																																																																																																																																																																																																																								
=.....								2																																																																																																																																																																																																																																																																								
								All																																																																																																																																																																																																																																																																								
Do not build a self-organizing vector of atoms pairs proportions	Construct a self-organizing vector of proportions of pairs of atoms for oxygen and nitrogen, as well as the ratio of the numbers of double and triple covalent bonds in molecules.	Construct a self-organizing vector of proportions of pairs of atoms for a list including atoms, chlorine, bromine, nitrogen, oxygen, Sulphur, phosphorus, as well as double and triple covalent bonds.																																																																																																																																																																																																																																																																														

Fig. 3.14 Examples of various variants of the self-organizing vector of atom pairs of proportions (APP)

Contributions	Contributions																		
<table border="1"> <tbody> <tr> <td><input type="checkbox"/> N</td> <td><input type="checkbox"/> O</td> <td><input type="checkbox"/> S</td> </tr> <tr> <td><input type="checkbox"/> F</td> <td><input type="checkbox"/> Cl</td> <td><input type="checkbox"/> Br</td> </tr> <tr> <td><input type="checkbox"/> =</td> <td><input type="checkbox"/> #</td> <td><input type="checkbox"/> Cmax</td> </tr> </tbody> </table>	<input type="checkbox"/> N	<input type="checkbox"/> O	<input type="checkbox"/> S	<input type="checkbox"/> F	<input type="checkbox"/> Cl	<input type="checkbox"/> Br	<input type="checkbox"/> =	<input type="checkbox"/> #	<input type="checkbox"/> Cmax	<table border="1"> <tbody> <tr> <td><input checked="" type="checkbox"/> N</td> <td><input type="checkbox"/> O</td> <td><input type="checkbox"/> S</td> </tr> <tr> <td><input type="checkbox"/> F</td> <td><input type="checkbox"/> Cl</td> <td><input type="checkbox"/> Br</td> </tr> <tr> <td><input type="checkbox"/> =</td> <td><input type="checkbox"/> #</td> <td><input checked="" type="checkbox"/> Cmax</td> </tr> </tbody> </table>	<input checked="" type="checkbox"/> N	<input type="checkbox"/> O	<input type="checkbox"/> S	<input type="checkbox"/> F	<input type="checkbox"/> Cl	<input type="checkbox"/> Br	<input type="checkbox"/> =	<input type="checkbox"/> #	<input checked="" type="checkbox"/> Cmax
<input type="checkbox"/> N	<input type="checkbox"/> O	<input type="checkbox"/> S																	
<input type="checkbox"/> F	<input type="checkbox"/> Cl	<input type="checkbox"/> Br																	
<input type="checkbox"/> =	<input type="checkbox"/> #	<input type="checkbox"/> Cmax																	
<input checked="" type="checkbox"/> N	<input type="checkbox"/> O	<input type="checkbox"/> S																	
<input type="checkbox"/> F	<input type="checkbox"/> Cl	<input type="checkbox"/> Br																	
<input type="checkbox"/> =	<input type="checkbox"/> #	<input checked="" type="checkbox"/> Cmax																	
No granting "extra powers" to any atom	"Additional weight" is given to the nitrogen atom and the number of rings in the molecule (Cmax)																		

Fig. 3.15 Possible uses for additional individual atom weights

$$\min(x, y) = \begin{cases} x, & \text{if } x > y \\ y, & \text{otherwise} \end{cases} \quad (3.6)$$

$$\max(x, y) = \begin{cases} x, & \text{if } x > y \\ y, & \text{otherwise} \end{cases} \quad (3.7)$$

$$^{-}\text{MAE} = \frac{1}{^{-}N} \sum |\Delta_k|, \quad ^{-}N \text{ is the number of } \Delta_k < 0 \quad (3.8)$$

$$^{+}\text{MAE} = \frac{1}{^{+}N} \sum |\Delta_k|, \quad ^{+}N \text{ is the number of } X_k \geq 0 \quad (3.9)$$

$$\Delta_k = \text{observed}_k - \text{calculated}_k \quad (3.10)$$

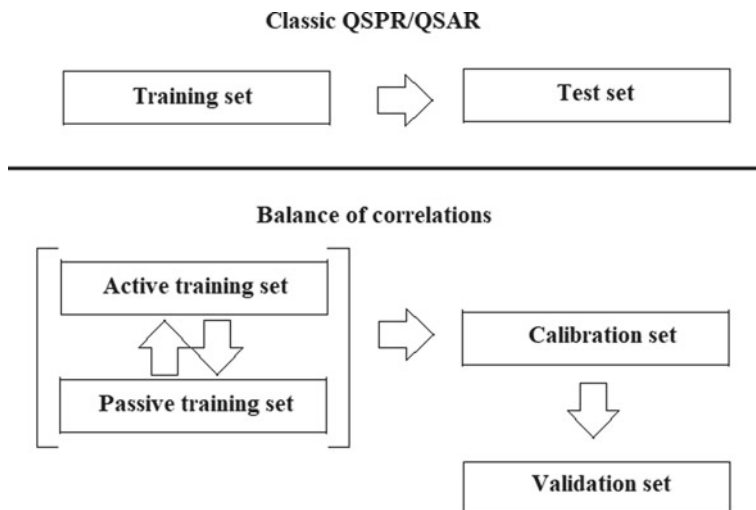


Fig. 3.16 Comparison of the classic scheme and the balance of correlations

The observed and calculated are corresponding values of the endpoint. Having data on all Δ_k for the calibration set, one can calculate the sum of negative ($-MAE$) and positive ($+MAE$) values of Δ_k , similar to traditional mean absolute error (MAE).

The CII is calculated as follows:

$$CII = 1 - \sum \text{Protest}_k \quad (3.11)$$

$$\text{Protest}_k = \begin{cases} R_k^2 - R^2, & \text{if } R_k^2 - R^2 > 0 \\ 0, & \text{otherwise} \end{cases} \quad (3.12)$$

The R^2 is the correlation coefficient for a set that contains n substances. The R_k^2 is the correlation coefficient for $n - 1$ substances of a set, after removing of k th substance. Hence, if the $(R_k^2 - R^2)$ is larger than zero, the k th substance is an “opponentist” for the correlation between experimental and predicted values of the set. A small sum of “protests” means a more “intensive” correlation.

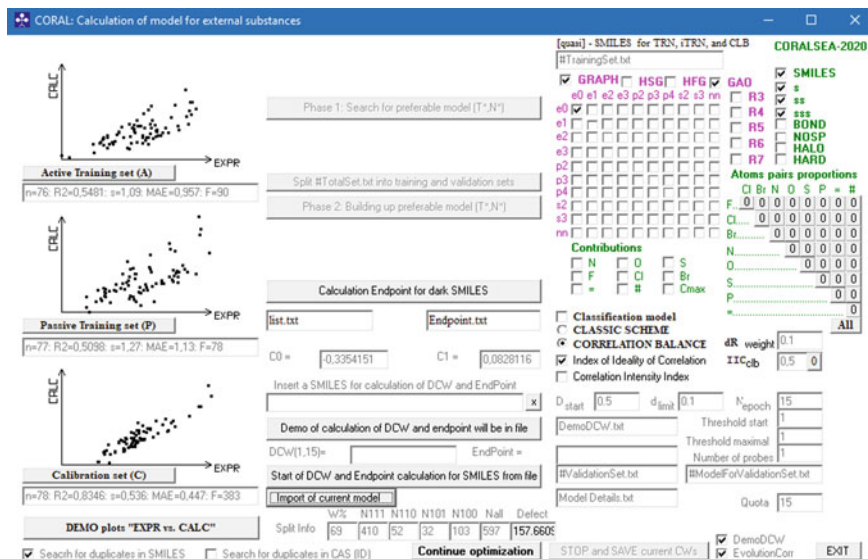


Fig. 3.17 An example of a model that is built up using the balance of correlations

3.1.2 An Example of Model Training and Validation (Graphically)

Figure 3.17 contains an example of a model for toxicity towards *Rainbow Trout* (LC_{50}) built using the balance of correlations. The interface provides a user with information on the selected method and the statistical quality of the model on all sets used to construct the model.

Figure 3.18 contains the results of applying the model for the external validation set and demonstrates a strange quality of the CORAL models calculated with IIC.

The active and passive training sets are divided into pairs of clusters (red colour shows calculated values that are overestimated, green colour shows calculated values that are underestimated). Thus, the involvement of IIC (as well CII) leads to an improvement in the quality of the model for the calibration set and for the validation set, but to the detriment of the statistical quality of the model for both training sets.

3.2 Conclusions

The SMILES concept has found numerous applications. Moreover, new SMILES applications are currently emerging for both applied and theoretical research in the field of physics, chemistry, and biology, as well as at the intersections of the natural sciences. SMILES modifications, both in practical and general theoretical terms, are

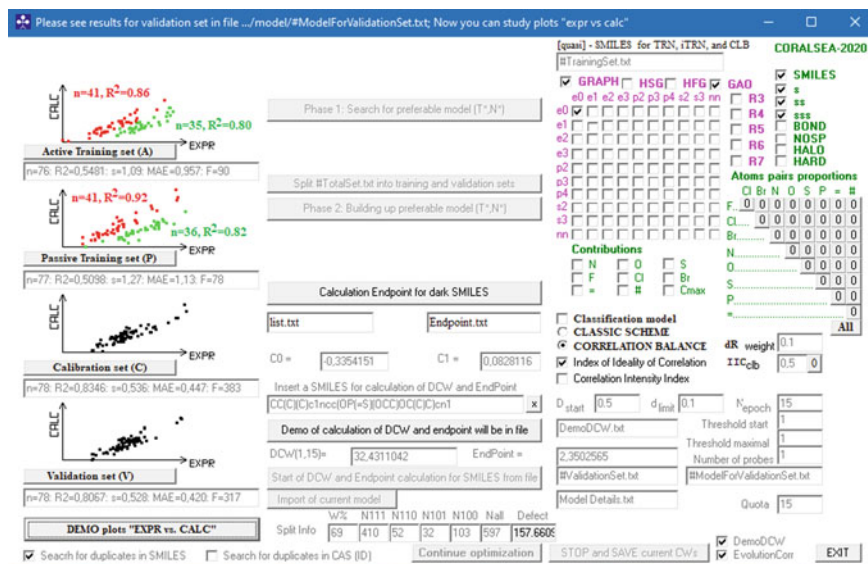


Fig. 3.18 Strange quality (indicated by red and green) of the CORAL models calculated with IIC and/or CII

also an important attribute of modern natural sciences. The CORAL program is one of the possible ways to use SMILES for building models of various endpoints, as well as for solving other problems in the field of natural sciences.

References

- Weininger D (1988) *J Chem Inf Comput Sci* 28:31–36. <https://doi.org/10.1021/ci00057a005>
- Weininger D, Weininger A, Weininger JL (1989) *J Chem Inf Comput Sci* 29:97–101. <https://doi.org/10.1021/ci00062a008>
- Weininger D (1990) *J Chem Inf Comput Sci* 30:237–243. <https://doi.org/10.1021/ci00067a005>
- Toropov AA, Toropova AP, Mukhamedzhanova DV, Gutman I (2005) *Indian J Chem Inorg Phys Theor Anal Chem* 44(8):1545–1552
- Toropov AA, Toropova AP, Benfenati E, Leszczynska D, Leszczynski J (2010) *J Comput Chem* 31(2):381–392. <https://doi.org/10.1002/jcc.21333>
- Enoch SJ, Madden JC, Cronin MTD (2008) *SAR QSAR Environ Res* 19(5–6):555–578. <https://doi.org/10.1080/10629360802348985>
- Toropov AA, Toropova AP, Benfenati E, Manganaro A (2009) *J Comput Chem* 30:2576–2582. <https://doi.org/10.1002/jcc.21263>
- Toropov AA, Toropova AP, Benfenati E, Leszczynska D, Leszczynski J (2009) *J Math Chem* 46(4):1232–1251. <https://doi.org/10.1007/s10910-008-9514-0>
- Toropova AP, Toropov AA, Benfenati E, Gini G (2011) *Chem Biol Drug Des* 77:343–360. <https://doi.org/10.1111/j.1747-0285.2011.01109.x>
- Toropov AA, Toropova AP, Benfenati E, Leszczynska D, Leszczynski J (2010) *Eur J Med Chem* 45:1387–1394. <https://doi.org/10.1016/j.ejmech.2009.12.037>

11. Toropov AA, Toropova AP, Benfenati E (2010) *Mol Divers* 14:183–192. <https://doi.org/10.1007/s11030-009-9156-6>
12. Toropov AA, Toropova AP, Benfenati E (2009) *J Math Chem* 46:1060–1073. <https://doi.org/10.1007/s10910-008-9491-3>
13. Toropov AA, Toropova AP, Benfenati E, Leszczynska D, Leszczynski J (2009) *J Math Chem* 47:355–369. <https://doi.org/10.1007/s10910-009-9574-9>
14. Olah M, Bologa C, Oprea TI (2004) *J Comput Aided Mol Des* 18(7–9):437–449. <https://doi.org/10.1007/s10822-004-4060-8>
15. Thalheim T, Vollmer A, Ebert R-U, Kühne R, Schüürmann G (2010) *J Chem Inf Model* 50(7):1223–1232. <https://doi.org/10.1021/ci1001179>
16. Ma EYT, Kremer SC (2009) In: *IEEE international conference on bioinformatics and biomedicine (BIBM)*, vol 5341870, pp 37–42. <https://doi.org/10.1109/BIBM.2009.60>
17. Saracci R (2006) *Int J Epidemiol* 35(3):513–514. <https://doi.org/10.1093/ije/dy1101>
18. Gutman I, Trinajstić N (1972) *Chem Phys Lett* 17(4):535–538. [https://doi.org/10.1016/0009-2614\(72\)85099-1](https://doi.org/10.1016/0009-2614(72)85099-1)
19. Randić M (1974) *J Chem Phys* 60:3920–3928. <https://doi.org/10.1063/1.1680839>
20. Gutman I, Ruščić B, Trinajstić N, Wilcox CF Jr (1975) *J Chem Phys* 62(9):3399–3405. <https://doi.org/10.1063/1.430994>
21. Balaban AT (1983) *Pure Appl Chem* 55(2):199–206. <https://doi.org/10.1351/pac198855020199>
22. Kier LB (1985) *Quant Struct-Act Relat* 4(3):109–116. <https://doi.org/10.1002/qsar.19850040303>
23. Mekenyan O, Bonchev D, Balaban A (1988) *J Math Chem* 2(4):347–375. <https://doi.org/10.1007/BF01166300>
24. Diudea MV (1994) *J Chem Inf Comput Sci* 34(5):1064–1071. <https://doi.org/10.1021/ci00021a005>
25. Basak SC, Bertelsen S, Grunwald GD (1994) *J Chem Inf Comput Sci* 34(2):270–276. <https://doi.org/10.1021/ci00018a007>
26. Bonchev D, Balaban AT, Liu X, Klein DJ (1994) *Int J Quantum Chem* 50(1):1–20. <https://doi.org/10.1002/qua.560500102>
27. Estrada E (1997) *J Chem Inf Comput Sci* 37(2):320–328. <https://doi.org/10.1021/ci960113v>
28. Ivanciuc O (2000) *J Chem Inf Comput Sci* 40(6):1412–1422. <https://doi.org/10.1021/ci00068y>
29. Pogliani L (2000) *Chem Rev* 100(10):3827–3858. <https://doi.org/10.1021/cr0004456>
30. Fujita S, Karasawa Y, Fujii M, Hironaka K-I, Uda S, Kubota H, Inoue H, Sumitomo Y, Hirayama A, Soga T, Kuroda S (2022) *NPJ Syst Biol Appl* 8(1):6. <https://doi.org/10.1038/s41540-022-00213-0>
31. Sheng C, Zhang W, Ji H, Zhang M, Song Y, Xu H, Zhu J, Miao Z, Jiang Q, Yao J, Zhou Y, Zhu J, Lü J (2006) *J Med Chem* 49(8):2512–2525. <https://doi.org/10.1021/jm051211n>
32. Klimisch H-J, Andrae M, Tillmann U (1997) *Regul Toxicol Pharmacol* 25(1):1–5. <https://doi.org/10.1006/rtp.1996.1076>
33. Taniguchi T, Tanaka K, Wang HO (2000) *IEEE Trans Fuzzy Syst* 8(4):442–452. <https://doi.org/10.1109/91.868950>
34. Guha R, Jurs PC (2005) *J Chem Inf Model* 45(3):800–806. <https://doi.org/10.1021/ci050022a>
35. Casañola-Martín GM, Marrero-Ponce Y, Khan MTH, Ather A, Sultan S, Torrens F, Rotondo R (2007) *Bioorg Med Chem* 15(3):1483–1503. <https://doi.org/10.1016/j.bmc.2006.10.067>
36. Tseng YJ, Hopfinger AJ, Esposito EX (2012) *J Comput Aided Mol Des* 26(1):39–43. <https://doi.org/10.1007/s10822-011-9511-4>
37. Parthasarathi R, Subramanian V, Roy DR, Chattaraj PK (2004) *Bioorg Med Chem* 12(21):5533–5543. <https://doi.org/10.1016/j.bmc.2004.08.013>
38. Hall LH, Kier LB (2001) *J Mol Graph Model* 20(1):4–18. [https://doi.org/10.1016/S1093-3263\(01\)00097-3](https://doi.org/10.1016/S1093-3263(01)00097-3)
39. Feng G-S (2012) *Cancer Cell* 21(2):150–154. <https://doi.org/10.1016/j.ccr.2012.01.001>

40. Joshi CP, Mansfield SD (2007) *Curr Opin Plant Biol* 10(3):220–226. <https://doi.org/10.1016/j.pbi.2007.04.013>
41. Karelson M, Maran U, Wang Y, Katritzky AR (1999) *Collect Czechoslov Chem Commun* 64(1):1551–1571. <https://doi.org/10.1135/cccc19991551>
42. Box GEP (1976) *J Am Stat Assoc* 71(356):791–799. <https://doi.org/10.1080/01621459.1976.10480949>
43. Box GEP, Tiao GC (1976) *J Appl Stat* 25(3):195–200. <https://doi.org/10.2307/2347226>
44. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T (2005) *ATLA Altern Lab Anim* 33(5):445–459. <https://doi.org/10.1177/026119290503300508>
45. Ellison CM, Sherhod R, Cronin MTD, Enoch SJ, Madden JC, Judson PN (2011) *J Chem Inf Model* 51(5):975–985. <https://doi.org/10.1021/ci1000967>
46. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R (2012) *Molecules* 17(5):4791–4810. <https://doi.org/10.3390/molecules17054791>
47. Li T, Zhao X, Li L (2021) *IEEE Trans Pattern Anal Mach Intell* (in press). <https://doi.org/10.1109/TPAMI.2021.3120428>
48. Gomari DP, Schweickart A, Cerchietti L, Paietta E, Fernandez H, Al-Amin H, Suhre K, Krumsiek J (2022) *Commun Biol* 5(1):645. <https://doi.org/10.1038/s42003-022-03579-3>
49. Fuhr AS, Sumpter BG (2022) *Front Mater* 9:865270. <https://doi.org/10.3389/fmats.2022.865270>
50. Sridharan B, Goel M, Priyakumar UD (2022) *ChemComm* 58(35):5316–5331. <https://doi.org/10.1039/d1cc07035e>
51. Monroe JJ, Shen VK (2022) *J Chem Theory Comput* 18(6):3622–3636. <https://doi.org/10.1021/acs.jctc.2c00110>
52. Aoto S, Hangai M, Ueno-Yokohata H, Ueda A, Igarashi M, Ito Y, Tsukamoto M, Jinno T, Sakamoto M, Okazaki Y, Hasegawa F, Ogata-Kawata H, Namura S, Kojima K, Kikuya M, Matsubara K, Taniguchi K, Okamura K (2022) *Sci Rep* 12(1):3730. <https://doi.org/10.1038/s41598-022-07560-2>
53. Kumar R, Sharma A, Alexiou A, Bilgrami AL, Kamal MA, Ashraf GM (2022) *Front Neurosci* 16:858126. <https://doi.org/10.3389/fnins.2022.858126>
54. Zhang X-C, Yi J-C, Yang G-P, Wu C-K, Hou T-J, Cao D-S (2022) *Brief Bioinform* 23(2):bbac033. <https://doi.org/10.1093/bib/bbac033>
55. Lim S, Lee YO, Yoon J, Kim YJ (2022) *J Comput-Aided Mol Des* 36(3):225–235. <https://doi.org/10.1007/s10822-022-00448-3>
56. Wang S, Liu J, Ding M, Gao Y, Liu D, Tian Q, Zhu J (2022) *Comb Chem High Throughput Screen* 25(4):642–650. <https://doi.org/10.2174/1386207324666210219102728>
57. Han X, Xie R, Li X, Li J (2022) *Life* 12(2):319. <https://doi.org/10.3390/life12020319>
58. Hung C, Gini G (2021) *Mol Divers* 25(3):1283–1299. <https://doi.org/10.1007/s11030-021-10250-2>
59. Probst D, Reymond J-L (2018) *J Chem Inf Model* 58(1):1–7. <https://doi.org/10.1021/acs.jcim.7b00425>
60. Přívratský J, Novák J (2021) *J Cheminform* 13(1):51. <https://doi.org/10.1186/s13321-021-00530-2>
61. Berenger F, Tsuda K (2021) *J Cheminform* 13(1):88. <https://doi.org/10.1186/s13321-021-00566-4>
62. Rajan K, Zielesny A, Steinbeck C (2020) *J Cheminform* 12(1):65. <https://doi.org/10.1186/s13321-020-00469-w>
63. Arús-Pous J, Johansson SV, Prykhodko O, Bjerrum EJ, Tyrchan C, Reymond J-L, Chen H, Engkvist O (2019) *J Cheminform* 11(1):71. <https://doi.org/10.1186/s13321-019-0393-0>
64. Drefahl A (2011) *J Cheminform* 3:1. <https://doi.org/10.1186/1758-2946-3-1>
65. Toropova AP, Toropov AA, Benfenati E, Leszczynska D, Leszczynski J (2018) *BioSystems* 169–170:5–12. <https://doi.org/10.1016/j.biosystems.2018.05.003>
66. Toropov AA, Toropova AP (2015) *Chemosphere* 139:18–22. <https://doi.org/10.1016/j.chemosphere.2015.05.042>

67. Toropova AP, Toropov AA, Manganelli S, Leone C, Baderna D, Benfenati E, Fanelli R (2016) *NanoImpact* 1:60–64. <https://doi.org/10.1016/j.impact.2016.04.003>
68. Toropova AP, Toropov AA (2022) *Sci Total Environ* 823:153747. <https://doi.org/10.1016/j.scitotenv.2022.153747>
69. Toropova AP, Toropov AA (2021) *Int J Environ Res* 15(4):709–722. <https://doi.org/10.1007/s41742-021-00346-w>
70. Toropov AA, Kjeldsen F, Toropova AP (2022) *Chemosphere* 303:135086. <https://doi.org/10.1016/j.chemosphere.2022.135086>
71. Weisgerber DW (1997) *J Am Soc Inf Sci* 48(4):349–360. [https://doi.org/10.1002/\(SICI\)1097-4571\(199704\)48:4%3c349::AID-ASI8%3e3.0.CO;2-W](https://doi.org/10.1002/(SICI)1097-4571(199704)48:4%3c349::AID-ASI8%3e3.0.CO;2-W)
72. O'Boyle NM (2012) *J Cheminform* 4(9):22. <https://doi.org/10.1186/1758-2946-4-22>
73. Österberg T, Norinder U (2001) *Eur J Pharm Sci* 12(3):327–337. [https://doi.org/10.1016/S0928-0987\(00\)00189-5](https://doi.org/10.1016/S0928-0987(00)00189-5)
74. Toropova AP, Toropov AA (2019) *J Mol Struct* 1182:141–149. <https://doi.org/10.1016/j.molstruc.2019.01.040>
75. Mukhopadhyay SK, Ahmad MO, Swamy MNS (2017) *Biomed Signal Process Control* 31:470–482. <https://doi.org/10.1016/j.bspc.2016.09.021>
76. Toropov A, Toropova A (2004) *J Mol Struct THEOCHEM* 711(1–3):173–183. <https://doi.org/10.1016/j.theochem.2004.10.003>
77. Toropova AP, Toropov AA, Benfenati E (2021) *Struct Chem* 32(3):967–971. <https://doi.org/10.1007/s11224-021-01778-y>
78. Toropov AA, Rasulev BF, Leszczynski J (2008) *Bioorg Med Chem* 16(11):5999–6008. <https://doi.org/10.1016/j.bmc.2008.04.055>
79. Achary PGR (2014) *SAR QSAR Environ Res* 25(6):507–526. <https://doi.org/10.1080/1062936X.2014.899267>
80. Toropov AA, Toropova AP, Benfenati E, Leszczynska D, Leszczynski J (2010) *Eur J Med Chem* 45(4):1387–1394. <https://doi.org/10.1016/j.ejmech.2009.12.037>
81. Kumar P, Kumar A (2018) *Drug Res* 68(4):189–195. <https://doi.org/10.1055/s-0043-119288>
82. Kumar P, Kumar A (2020) *Chemometr Intell Lab Syst* 200:103982. <https://doi.org/10.1016/j.chemolab.2020.103982>
83. Toropov AA, Toropova AP (2017) *Mutat Res Genet Toxicol Environ Mutagen* 819:31–37. <https://doi.org/10.1016/j.mrgentox.2017.05.008>
84. Toropova AP, Toropov AA (2017) *Sci Total Environ* 586:466–472. <https://doi.org/10.1016/j.scitotenv.2017.01.198>
85. Ahmadi S (2020) *Chemosphere* 242:125192. <https://doi.org/10.1016/j.chemosphere.2019.125192>
86. Kumar P, Kumar A, Sindhu J, Lal S (2019) *Drug Res* 69(3):159–167. <https://doi.org/10.1055/a-0652-5290>
87. Jain S, Amin SA, Adhikari N, Jha T, Gayen S (2020) *J Biomol Struct Dyn* 38(1):66–77. <https://doi.org/10.1080/07391102.2019.1566093>
88. Golubović M, Lazarević M, Zlatanović D, Krtinić D, Stoičkov V, Mladenović B, Milić DJ, Sokolović D, Veselinović AM (2018) *Comput Biol Chem* 75:32–38. <https://doi.org/10.1016/j.combiolchem.2018.04.009>
89. Duhhan M, Sindhu J, Kumar P, Devi M, Singh R, Kumar R, Lal S, Kumar A, Kumar S, Hussain K (2022) *J Biomol Struct Dyn* 40(11):4933–4953. <https://doi.org/10.1080/07391102.2020.1863861>
90. Toropov AA, Toropova AP (2020) *Sci Total Environ* 737:139720. <https://doi.org/10.1016/j.scitotenv.2020.139720>
91. Kumar P, Kumar A (2021) *J Mol Struct* 1246:131205. <https://doi.org/10.1016/j.molstruc.2021.131205>
92. Kumar P, Kumar A, Singh D (2022) *Environ Toxicol Pharm* 93:103893. <https://doi.org/10.1016/j.etap.2022.103893>

Part II
SMILES Based Descriptors

Chapter 4

All SMILES Variational Autoencoder for Molecular Property Prediction and Optimization



Zaccary Alperstein, Artem Cherkasov, and Jason Tyler Rolfe

Abstract Variational autoencoders (VAEs) defined over SMILES string and graph-based representations of molecules promise to improve the optimization of molecular properties, thereby revolutionizing the pharmaceuticals and materials industries. However, these VAEs are hindered by the non-unique nature of SMILES strings and the computational cost of graph convolutions. To efficiently pass messages along all paths through the molecular graph, we encode multiple SMILES strings of a single molecule using a set of stacked recurrent neural networks, harmonizing hidden representations of each atom between SMILES representations, and use attentional pooling to build a final fixed-length latent representation. By then decoding to a disjoint set of SMILES strings of the molecule, our All SMILES VAE learns an almost bijective mapping between molecules and latent representations near the high probability mass subspace of the prior. Our SMILES-derived but molecule-based latent representations significantly surpass the state of the art in a variety of fully and semi-supervised property regression and molecular property optimization tasks.

Keywords SMILES · Variational autoencoders · Optimization · Validation

This work was performed while Zaccary Alperstein and Jason Rolfe were working at D-Wave Systems. The work herein, the research behind it, and all associated material is copyright ©D-Wave and UBC.

Z. Alperstein · J. T. Rolfe (✉)
Variational AI, 210-577 Great Northern Way, Vancouver, BC V5T 1E1, Canada
e-mail: jason@variational.ai

Z. Alperstein
e-mail: zac@variational.ai

A. Cherkasov
Vancouver Prostate Centre, UBC, 2660 Oak Street, Vancouver, BC V6H 3Z6, Canada
e-mail: acherkasov@prostatecentre.com

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. P. Toropova and A. A. Toropov (eds.), *QSPR/QSAR Analysis Using SMILES and Quasi-SMILES*, Challenges and Advances in Computational Chemistry and Physics 33, https://doi.org/10.1007/978-3-031-28401-4_4

4.1 Introduction

The design of new pharmaceuticals, OLED materials, and photovoltaics all requires optimization within the space of molecules [1]. While well-known algorithms ranging from gradient descent to the simplex method facilitate efficient optimization, they generally assume a continuous search space and a smooth objective function. In contrast, the space of molecules is discrete and sparse. Molecules correspond to graphs, with each node labeled by one of ninety-eight naturally occurring atoms, and each edge labeled as a single, double, or triple bond. Even within this discrete space, almost all possible combinations of atoms and bonds do not form chemically stable molecules, and so must be excluded from the optimization domain, yet there remain as many as 10^{60} small molecules to consider [2]. Moreover, properties of interest are often sensitive to even small changes to the molecule [3], so their optimization is intrinsically difficult.

Efficient, gradient-based optimization can be performed over the space of molecules given a map between a continuous space, such as \mathbb{R}^n or the n -sphere, and the space of molecules and their properties [4]. Initial approaches of this form trained a variational autoencoder (VAE) [5, 6] on SMILES string representations of molecules [7] to learn a decoder mapping from a Gaussian prior to the space of SMILES strings [8]. A sparse Gaussian process on molecular properties then facilitates Bayesian optimization of molecular properties within the latent space [8–11], or a neural network regressor from the latent space to molecular properties can be used to perform gradient descent on molecular properties with respect to the latent space [12–15]. Alternatively, semi-supervised VAEs condition the decoder on the molecular properties [16, 17], so the desired properties can be specified directly. Recurrent neural networks have also been trained to model SMILES strings directly and tuned with transfer learning, without an explicit latent space or encoder [18, 19].

SMILES, the simplified molecular-input line-entry system, defines a character string representation of a molecule by performing a depth-first pre-order traversal of a spanning tree of the molecular graph, emitting characters for each atom, bond, tree-traversal decision, and broken cycle [7]. The resulting character string corresponds to a flattening of a spanning tree of the molecular graph, as shown in Fig. 4.1. The SMILES grammar is restrictive, and most strings over the appropriate character set do not correspond to well-defined molecules. Rather than require the VAE decoder to explicitly learn this grammar, context-free grammars [10] and attribute grammars [9] have been used to constrain the decoder, increasing the percentage of valid SMILES strings produced by the generative model. Invalid SMILES strings and violations of simple chemical rules can be avoided entirely by operating on the space of molecular graphs, either directly [14, 20–23] or via junction trees [13].

Every molecule is represented by many well-formed SMILES strings, corresponding to all depth-first traversals of every spanning tree of the molecular graph. The distance between different SMILES strings of the same molecule can be much greater than that between SMILES strings from radically dissimilar molecules [13], as shown in Fig. 4.2. A generative model of individual SMILES strings will tend to reflect this

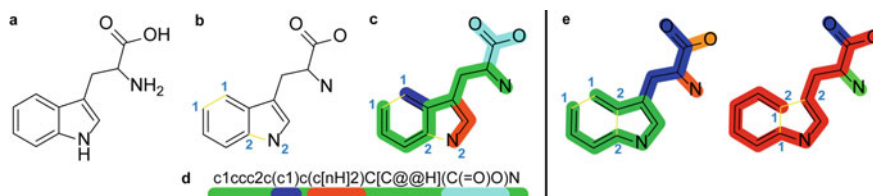


Fig. 4.1 Molecular graph of the amino acid Tryptophan (a). To construct a SMILES string, all cycles are broken, forming a spanning tree (b); a depth-first traversal is selected (c); and this traversal is flattened (d). The beginning and end of intermediate branches in the traversal are denoted by "(" and ")" respectively. The ends of broken cycles are indicated with matching digits. The full grammar is listed in Sect. 4.7. A small set of SMILES strings can cover all paths through a molecule (e)

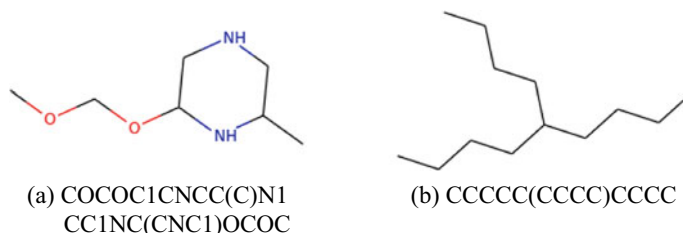


Fig. 4.2 Multiple SMILES strings of a single molecule may be more dissimilar than SMILES strings of radically dissimilar molecules. The top SMILES string for molecule (a) is 30% similar to the bottom SMILES string by string edit distance, but 60% similar to the SMILES string for molecule (b)

geometry, complicating the mapping from latent space to molecular properties and creating unnecessary local optima for property optimization [24]. To address this difficulty, sequence-to-sequence transcoders [25] have been trained to map between different SMILES strings of a single molecule [26–29].

Reinforcement learning, often combined with adversarial methods, has been used to train progressive molecule growth strategies [30–35]. While these approaches have achieved state-of-the-art optimization of simple molecular properties that can be evaluated quickly *in silico*, critic-free techniques generally depend upon property values of algorithm-generated molecules (but see [20, 36]) and so scale poorly to real-world properties requiring time-consuming wet laboratory experiments.

Molecular property optimization would benefit from a generative model that directly captures the geometry of the space of molecular graphs, rather than SMILES strings, but efficiently infers a latent representation sensitive to spatially distributed molecular features. To this end, we introduce the All SMILES VAE, which uses recurrent neural networks (RNNs) on multiple SMILES strings to implicitly perform efficient message passing along and among many flattened spanning trees of the molecular graph in parallel. A fixed-length latent representation is distilled from the variable-length RNN output using attentional mechanisms. From this latent representation, the decoder RNN reconstructs a set of SMILES strings disjoint from

those input to the encoder, ensuring that the latent representation only captures features of the molecule, rather than its SMILES realization. Simple property regressors jointly trained on this latent representation surpass the state of the art for molecular property prediction and facilitate exceptional gradient-based molecular property optimization when constrained to the region of the prior containing almost all the probability around it. We further demonstrate that the latent representation forms a near bijection with the space of molecules and is smooth with respect to molecular properties, facilitating effective optimization.

4.1.1 *Summary of Novel Contributions*

Starting with the work of Gómez-Bombarelli et al. [8], previous molecular variational autoencoders have used one consistent SMILES string as both the input to the RNN encoder and the target of the RNN decoder. Any single SMILES string explicitly represents only a subset of the pathways in the molecular graph. Correspondingly, the recurrent neural networks in these encoders implicitly propagated information through only a fraction of the possible pathways. Kipf and Welling [37], Liu et al. [14], and Simonovsky and Komodakis [23], among others, trained molecular VAEs with graph convolutional encoders, which pass information through all graph pathways in parallel, but at considerable computational expense. None of these works used enough layers of graph convolutions to transfer information across the diameter of the average molecule in standard drug design datasets. This is partially overcome by Lusci et al. [38] who ensemble RNN-based representations of multiple directed-acyclic graphs of a single molecule for property prediction. The All SMILES VAE introduces the use of multiple SMILES strings of a single, common molecule as input to a RNN encoder, with pooling of homologous messages among the hidden representations associated with different SMILES strings. This allows information to flow through all pathways of the molecular graph, but can efficiently propagate information across the entire width of the molecule in a single layer.

Bjerrum and Sattarov [27] and Winter et al. [29] trained sequence-to-sequence transcoders to map between different SMILES strings of the same molecule. These transcoders do not define an explicit generative model over molecules, and their latent spaces have no prior distributions. The All SMILES VAE extends this approach to variational autoencoders and thereby learns a SMILES-derived generative model of molecules, rather than SMILES strings. The powerful, learned, hierarchical prior of the All SMILES VAE regularizes molecular optimization and property prediction. To ensure that molecular property optimization searches within the practical support of the prior, containing almost all of its probability mass, we introduce a hierarchical radius constraint on optimization with respect to the latent space.

4.2 Efficient Molecular Encoding with Multiple SMILES Strings

A variational autoencoder (VAE) defines a generative model over an observed space x in terms of a prior distribution over a latent space $p(z)$ and a conditional likelihood of observed states given the latent configuration $p(x|z)$ [5, 6]. The true log-likelihood $\log [p(x)] = \log [\int_z p(z)p(x|z)]$ is intractable, so the evidence lower bound (ELBO), based upon a variational approximation $q(z|x)$ to the posterior distribution, is maximized instead:

$$\mathcal{L} = \mathbb{E}_{q(z|x)} [\log p(x|z)] - \text{KL} [q(z|x)||p(z)]. \quad (4.1)$$

The ELBO implicitly defines a stochastic autoencoder, with encoder $q(z|x)$ and decoder $p(x|z)$.

Many effective encoders for molecules rely upon graph convolutions: local message passing in the molecular graph, between either adjacent nodes or adjacent edges [38–42]. To maintain permutation symmetry, the signal into each node is a sum of messages from the adjacent nodes, but may be a function of edge type, or attentional mechanisms dependent upon the source and destination nodes [43]. This sum of messages is then subject to a linear transformation and a pointwise nonlinearity. Messages are sometimes subject to gating [42], like in long short-term memories (LSTM) [44] and gated recurrent units (GRU) [45], as detailed in Sect. 4.3.

More specifically, graph convolutions are conventionally defined by:

$$h_t^{(n)} = f \left(\left(\sum_{m \in \mathcal{N}(n)} h_{t-1}^{(m)} \right) W_t \right) \quad (4.2)$$

where $\mathcal{N}(n)$ is the set of neighbors of node n , for which there is an edge between n and $m \in \mathcal{N}(n)$, and $f(x)$ is a pointwise nonlinearity such as a logistic function or rectified linear unit. This message passing can be understood as a first-order approximation to spectral convolutions on graphs [46]. Kipf and Welling [41] additionally normalize each message by the square root of the degree of each node before and after the sum over neighboring nodes. Kearnes et al. [40] maintain separate messages for nodes and edges, with the neighborhood of a node comprising the connected edges and the neighborhood of an edge comprising the connected nodes. Li et al. [42] add gating analogous to a GRU.

Message passing on molecular graphs is analogous to a traditional convolutional neural network applied to images [47, 48], with constant-resolution hidden layers [49] and two kernels: a 3×3 average-pooling kernel that sums messages from adjacent pixels (corresponding to adjacent nodes in a molecular graph) and a trainable 1×1 kernel that transforms the message from each pixel (node) independently, before a pointwise nonlinearity. While convolutional networks with such small kernels are now standard in the visual domain, they use hundreds of layers to pass

information throughout the image and achieve effective receptive fields that span the entire input [50]. In contrast, molecule encoders generally use between three and seven rounds of message passing [11, 13, 14, 34, 39, 40, 51]. This limits the computational cost, since molecule encoders cannot use highly optimized implementations of spatial 2D convolutions, but each iteration of message passing only propagates information a geodesic distance of one within the molecular graph.¹ In the case of the commonly used ZINC250k dataset of 250,000 drug-like molecules [8], information cannot traverse these graphs effectively, as their average diameter is 11.1 and their maximum diameter is 24, as shown in Sect. 4.5.

Non-local molecular properties, requiring long-range information propagation along the molecular graph, are of practical interest in domains including pharmaceuticals, photovoltaics, and OLEDs. The pharmacological efficacy of a molecule generally depends upon high binding affinity for a particular receptor or other target, and low binding affinity for other possible targets. These binding affinities are determined by the maximum achievable alignment between the molecule’s electromagnetic fields and those of the receptor. Changes to the shape or charge distribution in one part of the molecule affect the position and orientation at which it fits best with the receptor, inducing shifts and rotations that alter the binding of other parts of the molecule and changing the binding affinity [52]. Similarly, efficient next-generation OLEDs depend on properties, such as the singlet-triplet energy gap, that are directly proportional to the strength of long-range electronic interactions across the molecule [53]. The latent representation of a VAE can directly capture these non-local, nonlinear properties only if the encoder passes information efficiently across the entire molecular graph.

Analogous to graph convolutions, gated RNNs defined directly on SMILES strings effectively pass messages, via the hidden state, through a flattened spanning tree of the molecular graph (see Fig. 4.1). The message at each symbol in the string is a weighted sum of the previous message and the current input, followed by a pointwise nonlinearity and subject to gating. This differs from explicit graph-based message passing in that the molecular graph is flattened into a chain corresponding to a depth-first pre-order traversal of a spanning tree, and the set of adjacent nodes that affect a message only includes the preceding node in this chain. Rather than updating all messages in parallel, RNNs on SMILES strings move sequentially down the chain, so earlier messages influence all later messages, and information can propagate through all branches of a flattening of a spanning tree in a single pass. With a well-chosen spanning tree, information can pass the entire width of the molecular graph in a single RNN update.

¹ All-to-all connections allow fast information transfer, but computation is quadratic in graph size [40, 51]. Lusci et al. [38] considered a set of DAGs rooted at every atom, with full message propagation in a single pass.

4.3 Review of Recurrent Neural Networks

Recurrent neural networks, such as long short-term memories (LSTMs) [44] and gated recurrent units (GRUs) [45], are commonly used to model text, audio, and other one-dimensional sequences. Gated recurrent units (GRUs) are defined by Cho et al. [45]:

$$\begin{aligned} [r, z] &= \sigma(x_t [W_r, W_z] + h_{t-1} [U_r, U_z] + [b_r, b_z]) \\ h_t &= (1 - z) \odot h_{t-1} + z \odot \tanh(x_t W + (r \odot h_{t-1}) U + b_h) \end{aligned}$$

where r , z , and h are row vectors, $[x, y]$ denotes the column-wise concatenation of x and y , and the logistic function $\sigma(x) = (1 + e^{-x})^{-1}$ and hyperbolic tangent are applied element-wise to vector argument x . The hidden state h_t , comprising the message from node t , is a gated, weighted sum of the previous message h_{t-1} and the current input x_t , both subject to an element-wise linear transformation and nonlinear (sigmoid) transformation. Specifically, the sum of the message from the input, $x_t W U^{-1}$ and the gated message from the previous node, $r \odot h_{t-1}$, is subject to a linear transformation U and a pointwise nonlinearity. This is then gated and added to a gated residual connection from the previous node.

Long short-term memories (LSTMs) are defined similarly [44]:

$$\begin{aligned} [f_t, i_t, o_t] &= \sigma(x_t [W_f, W_i, W_o] + h_{t-1} [U_f, U_i, U_o] + [b_f, b_i, b_o]) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(x_t W_c + h_{t-1} U_c + b_c) \\ h_t &= o_t \odot \tanh(c_t) \end{aligned}$$

where f is the forget gate, i is the input gate, and o is the output gate. LSTMs impose a second hyperbolic tangent and gating unit on the nonlinear recurrent message, but nevertheless still follow the form of applying width-two kernels and pointwise nonlinearities to the input and hidden state.

An LSTM, taking a SMILES string as input, can realize a subset of the messages passed by graph convolutions. For instance, input gates and forget gates can conspire to ignore open parentheses, which indicate the beginning of a branch of the depth-first spanning tree traversal. If they similarly ignore the digits that close broken rings, the messages along each branch of the flattened spanning tree are not affected by the extraneous SMILES syntax. Input and forget gates can then reset the LSTM's memory at close parentheses, which indicate the end of a branch of the depth-first spanning tree traversal, and the return to a previous node, ensuring that messages only propagate along connected paths in the molecular graph. While an LSTM decoder generating SMILES strings faces ambiguity regarding which of the set of SMILES strings representing a molecule to produce, this is analogous to the problem faced by graph-based decoders, as discussed in Sect. 4.7.2.

4.4 All SMILES VAE Architecture

To marry the latent space geometry induced by graph convolutions to the information propagation efficiency of RNNs on SMILES strings, the All SMILES encoder combines these architectures. It takes multiple distinct SMILES strings of the same molecule as input and applies RNNs to them in parallel. This implicitly realizes a representative set of message passing pathways through the molecular graph, corresponding to the depth-first pre-order traversals of the spanning trees underlying the SMILES strings. Between each layer of RNNs, the encoder harmonizes homologous messages between parallel representations of the multiple SMILES strings. In this harmonization, all messages to a single atom across the multiple SMILES strings are replaced with their pooled average, so that information flows along the union of the implicit SMILES pathways.

Initially, the characters of the multiple SMILES strings are linearly embedded, and each string is preprocessed by a bidirectional GRU (BiGRU) [45], followed by a linear transformation, to produce the layer 0 representation \mathbf{H}_i^0 for each SMILES string i . For each SMILES string i and layer l , \mathbf{H}_i^l is a sequence of vector embeddings, one for each character of the original SMILES string, collectively forming a matrix. The encoder then applies a stack of modules, each of which harmonizes atom representations across SMILES strings, followed by layer norm [54], concatenation with the linearly embedded SMILES input, and a GRU applied to the parallel representations independently, as shown in Figs. 4.3 and 4.4.

Multiple SMILES strings representing a single molecule need not have the same length, and syntactic characters indicating branching and ring closures rather than atoms and bonds do not generally match. However, the set of atoms is always consistent, and a bijection can be defined between homologous atom characters. At the beginning of each encoder module (Fig. 4.3), the parallel inputs corresponding to a single, common atom of the original molecule are pooled, as shown in Fig. 4.4. This harmonized atom representation replaces the original in each of the input streams for the subsequent layer normalizations and GRUs, reversing the information flow of Fig. 4.4. To realize atom harmonization, we experimented with average and max pooling, but found element-wise sigmoid gating to be most effective [42,

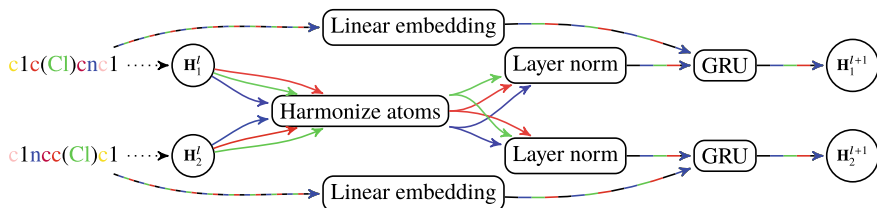


Fig. 4.3 In each layer of the encoder after the initial BiGRU and linear transformation, hidden states corresponding to each atom are harmonized across encodings of different SMILES strings for a common molecule, followed by layer norm and a GRU on each SMILES encoding independently

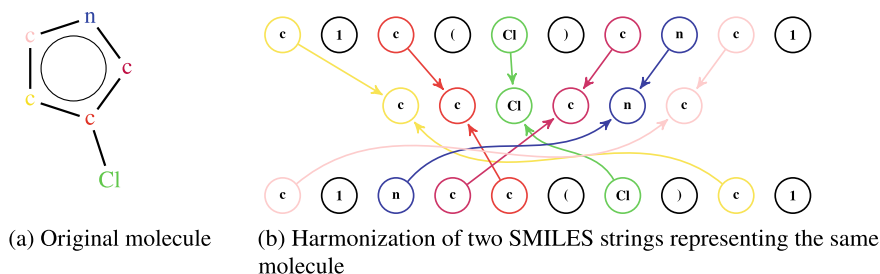


Fig. 4.4 To pass information between multiple SMILES representations of a molecule (a), the encoder harmonizes the representation of each atom. Homologous messages corresponding to the same atom are pooled (b), and the original messages are replaced with this pooled message, reversing the information flow of (b)

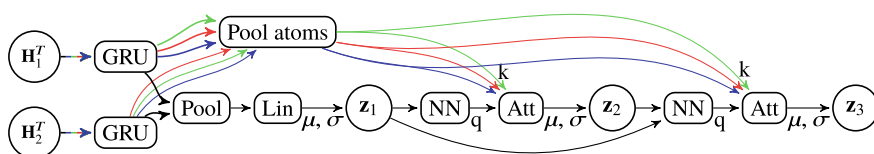


Fig. 4.5 The approximating posterior is an autoregressive set of Gaussian distributions. The mean (μ) and log-variance ($\log \sigma^2$) of the first subset of latent variables z_1 are a linear transformation of the max-pooled final hidden state of GRUs fed the encoder outputs. Succeeding subsets z_i are produced via Bahdanau-style attention with the pooled atom outputs of the GRUs as keys (k), and the query (q) computed by a neural network on $z_{<i}$

43, 55]: $a' = \frac{1}{k} \sum_k (a_k \odot \sigma(W[a_k, \frac{1}{k} \sum_k a_k] + b))$, where $[x, y]$ is the concatenation of vectors x and y and the logistic function $\sigma(x)$ is applied element-wise. The pooling effectively sums messages propagated from many adjacent nodes in the molecular graph, analogous to a graph convolution, but the GRUs efficiently transfer information through many edges in each layer, rather than just one. The hidden representations associated with non-atom, syntactic input characters, such as parentheses and digits, are left unchanged by the harmonization operation.

The approximating posterior distills the resulting variable-length encodings into a fixed-length hierarchy of autoregressive Gaussian distributions [56]. The mean and log-variance of the first layer of the approximating posterior, z_1 , are parametrized by max-pooling the terminal hidden states of the final encoder GRUs, followed by batch renormalization [57] and a linear transformation, as shown in Fig. 4.5.

Succeeding hierarchical layers use Bahdanau-style attention [58] over the pooled final atom vectors. Specifically, the final encoder hidden vectors for each atom comprise the key vectors k , whereas the query vector q is computed by a one-hidden layer network of rectified linear units given the concatenation of the previous latent layers as input. The final output of the attentional mechanism, c , is computed via:

$$\begin{aligned}
 e_i &= \tanh(qW_a + k_i U_a) v^\top \\
 \alpha_i &= \frac{\exp(e_i)}{\sum_j \exp(e_j)} \\
 c &= \sum_i \alpha_i k_i
 \end{aligned}$$

The output of the attentional mechanism is subject to batch renormalization and a linear transformation to compute the conditional mean and log-variance of the layer.

This is analogous to the order-invariant encoding of set2set, but an output is produced at each step, and processing is not gated [24]. The attentional mechanism is also effectively available to property regressors that take the fixed-length latent representation as input, allowing them to aggregate contributions from across the molecule. The prior has a similar autoregressive structure, but uses neural networks of ReLUs in place of Bahdanau-style attention, since it does not have access to the atom vectors. For molecular optimization tasks, we usually scale up the term $\text{KL}[q(z|x)||p(z)]$ in the ELBO by the number of SMILES strings in the decoder, analogous to multiple single-SMILES VAEs in parallel; we leave this KL term unscaled for property prediction.

The decoder is a single-layer LSTM, for which the initial cell state is computed from the latent representation $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2, \dots]$ by a neural network, and a linear transformation of the latent representation is concatenated onto each input. It is trained with teacher forcing to reconstruct a set of SMILES strings disjoint from those provided to the encoder, but representing the same molecule. As in conventional language models, the decoder LSTM autoregressively produces a sequence of categorical distributions for each successive SMILES character conditioned on the preceding characters. Grammatical constraints [9, 10] can naturally be enforced within this LSTM by parsing the unfolding character sequence with a pushdown automaton and constraining the final softmax of the LSTM output at each time step to grammatically valid symbols. This is detailed in Sect. 4.7, although we leave the exploration of this technique to future work.

The full All SMILES VAE architecture is summarized in Fig. 4.6. The evidence lower bound (ELBO) of the log-likelihood (Eq. 4.1) is the sum of the conditional log-likelihoods of \mathbf{x}'_i in Fig. 4.6, minus the Kullback–Leibler divergence between the approximating posterior, $q(z|x)$, computed by node AP in Fig. 4.6, and the prior depicted in Fig. 4.7.

The All SMILES VAE is a generative model over both the structure and properties of molecules \mathcal{M} , so we define the conditional likelihood to be

$$p(\mathcal{M}|z) = p(\rho^{\mathcal{M}}|z) \cdot \prod_j p(x_j^{\mathcal{M}}|z),$$

where $\{x_j^{\mathcal{M}}\}_{j=1}^N$ is a set of N SMILES strings of a molecule \mathcal{M} with properties $\rho^{\mathcal{M}}$. Unlike a conventional VAE, the representation of the molecule \mathcal{M} input to the encoder

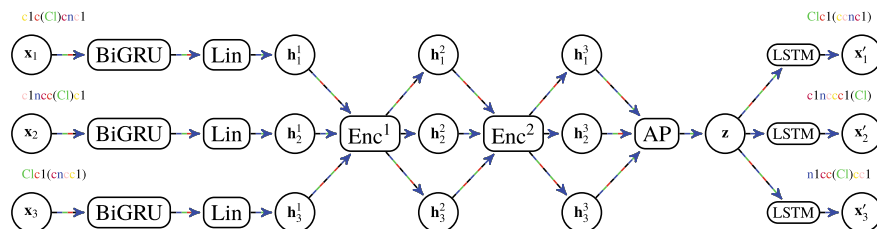


Fig. 4.6 Multiple SMILES strings representing a single, common molecule are preprocessed by a BiGRU and a linear transformation, followed by multiple encoder blocks as in Figs. 4.3 and 4.4. The approximating posterior depicted in Fig. 4.5 then produces a sample from the latent state \mathbf{z} , which is decoded into SMILES strings by LSTMs. Note that all SMILES strings, in both the input and the output, are distinct. The encoder blocks also receive a linear embedding of the original SMILES strings as input

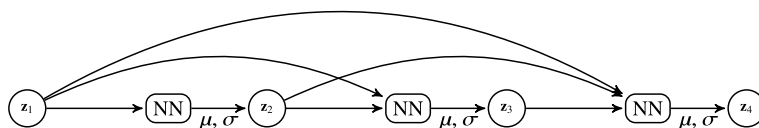


Fig. 4.7 Prior distribution over $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2, \dots]$ is a hierarchy of autoregressive Gaussians. The conditional prior distribution of hierarchical layer i given layers 1 through $i - 1$, $p(\mathbf{z}_i | \mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_{i-1})$, is a Gaussian with mean μ and log-variance $\log \sigma^2$ determined by a neural network with input $[\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_{i-1}]$

$q(\mathbf{z} | \mathcal{M})$ is not identical to the target of the conditional likelihood $p(\mathcal{M} | \mathbf{z})$; rather, it comprises a set of SMILES strings $\{x_i^{\mathcal{M}}\}_{i=1}^M$ of the molecule \mathcal{M} disjoint from the decoding target and does not include the molecular properties. Nevertheless, both encoder input and decoder target correspond to a single molecule \mathcal{M} . The conditional log-likelihood of the molecular properties $\log p(\rho^{\mathcal{M}} | \mathbf{z})$ is implicitly parametrized by scaling its contribution to the ELBO by λ . For instance, if $p(\rho^{\mathcal{M}} | \mathbf{z})$ is a unit-variance Gaussian distribution, then λ sets the effective variance to λ^{-1} . Finally, when optimizing molecular properties, we scale the KL term by M , the number of SMILES strings in the decoder, rendering the ELBO analogous to multiple single-SMILES VAEs in parallel. The resulting ELBO is

$$\mathcal{L} = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} | \{x_i\}_{i=1}^N)} \left[\lambda \cdot \log p(\rho | \mathbf{z}) + \sum_{j=1}^M \log p(x_j | \mathbf{z}) \right] - M \cdot \text{KL} \left[q(\mathbf{z} | \{x_i\}_{i=1}^N) \parallel p_{\theta}(\mathbf{z}) \right].$$

Since the SMILES inputs to the encoder are different from the targets of the decoder, the decoder is effectively trained to assign high probability to all SMILES strings of the encoded molecule. The latent representation must capture the molecule as a whole, rather than any particular SMILES input to the encoder. To accommodate this intentionally difficult reconstruction task, facilitate the construction of a bijection

between latent space and molecules, and following prior work [16, 29], we use a width-five beam search decoder to map from the latent representation to the space of molecules at test time.

In all experiments, we use a set of $M = 5$ randomly selected SMILES strings for encoding and $N = 5$ disjoint SMILES strings as the decoding target. We use encoder stacks of depth three, with 512 hidden units in each GRU. The approximating posterior uses four layers of hierarchy, with 128 hidden units in the one-hidden layer neural network that computes the attentional query vector. In practice, separate GRUs were used to produce the final hidden state for \mathbf{z}_1 and the atom representations for $\mathbf{z}_{>1}$. The single-layer LSTM decoder has 2048 hidden units. Training was performed using ADAM, with a decaying learning rate and KL annealing. In all multiple SMILES strings architectures, we use five SMILES strings for encoding and decoding which are selected with RDKit [59].

In contrast to many previous molecular VAEs, we do not scale down the term $\text{KL}[q(z|x)||p(z)]$ in the ELBO by the number of latent units [9, 10]. However, our loss function does include separate reconstructions for multiple SMILES strings of a single molecule. For molecular optimization tasks, we usually scale up this KL term by the number of SMILES strings in the decoder, analogous to multiple single-SMILES VAEs in parallel; we leave the KL term unscaled for property prediction.

4.4.1 Computational Complexity

Since the length of a SMILES string is linear in the total number of bonds b , the computational complexity of each layer of the All SMILES encoder is $\mathcal{O}(M \cdot b)$, where $M = 5$ is the number of random SMILES strings of the molecule. Similarly, the complexity of each layer of graph convolution is $\mathcal{O}(b)$. However, to pass information through the entire molecule, graph convolutions require a number of layers proportional to the graph diameter. Molecular graph convolutions generally use a fixed architecture for all molecules. In principle, the maximum diameter of a molecule is equal to the number of bonds. As a result, the computational complexity for graph convolutions to pass information through all molecules is $\mathcal{O}(b^2)$. In contrast, each RNN in the All SMILES encoder can in principle pass information through the entire graph, so the computational complexity remains $\mathcal{O}(M \cdot b)$.

4.4.2 Latent Space Optimization

Unlike many models that apply a sparse Gaussian process to fixed latent representations to predict molecular properties [9–11, 13], the All SMILES VAE jointly trains property regressors with the generative model (as do [14]).² We use linear regressors

² Gómez-Bombarelli, et al. [8] jointly train a regressor, but still optimize using a Gaussian process.

for the log octanol-water partition coefficient ($\log P$) and molecular weight (MW), which have unbounded values, and logistic regressors for the quantitative estimate of drug-likeness (QED) [60] and twelve binary measures of toxicity [61, 62], which take values in $[0, 1]$. We then perform gradient-based optimization of the property of interest with respect to the latent space and decode the result to produce an optimized molecule.

Naively, we might either optimize the predicted property without constraints on the latent space, or find the maximum a posteriori (MAP) latent point for a conditional likelihood over the property that assigns greater probability to more desirable values. However, the property regressors and decoder are only accurate within the domain in which they have been trained: the region assigned high probability mass by the prior. For a n -dimensional standard Gaussian prior, almost all probability mass lies in a practical support comprising a thin spherical shell of radius $\sqrt{n-1}$ [63]. With linear or logistic regressors, predicted property values increase monotonically in the direction of the weight vector, so unconstrained property maximization diverges from the origin of the latent space. Conversely, MAP optimization with a Gaussian prior is pulled toward the origin, where the density of the prior is greatest. Both unconstrained and MAP optimization thus deviate from the practical support in each layer of the hierarchical prior, resulting in large prediction errors and poor optimization.

We can use the reparametrization trick [5, 6] to map our autoregressive prior back to a standard Gaussian. The image of the thin spherical shell through this reparametrization still contains almost all of the probability mass. We therefore constrain optimization to the reparametrized $n-1$ dimensional sphere of radius $\sqrt{n-1}$ for each n -dimensional layer of the hierarchical prior by optimizing the angle directly.³ Although the reparametrization from the standard Gaussian prior to our autoregressive prior is not volume preserving, this hierarchical radius constraint holds us to the center of the image of the thin spherical shell. The distance to which the image of the thin spherical shell extends away from the $n-1$ dimensional sphere at its center is a highly nonlinear function of the previous layers.

The pseudocode for optimization in the latent space is shown in Algorithms 1 and 2. We project each layer of latent variables separately onto the radius defined by their conditional Gaussian distribution and then optimize with respect to the $n-1$ angles.

To further ensure that the optimization is constrained to well-trained regions of latent space, we add $\beta \cdot \log p(z)$ to the objective function, where β is a hyperparameter. Finally, to moderate the strictly monotonic nature of linear regressors, we apply an element-wise hard tanh to all latent variables before the regressor, with a linear region that encompasses all values observed in the training set.

To compare with previous work as fairly as possible, we optimize 1000 random samples from the prior to convergence, collecting the last point from each trajectory with a valid SMILES decoding. From these 1000 points, we evaluate the true molecular property on the 100 points for which the predicted property value is the largest. Of these 100 values, we report the three largest. However, optimization within our latent

³ This generalizes the slerp interpolations of Gómez-Bombarelli et al. [8] to optimization.

space is computationally inexpensive and requires no additional property measurement data. We could somewhat improve molecular optimization at minimal expense by constructing additional optimization trajectories in latent space and evaluating the true molecular properties on the best points from this larger set.

Algorithm 1: Initialize Angles

```

output: Angular coordinates of latent variable sample on the spherical shell with radius
           $\sqrt{N-1}$ 
for  $i \leftarrow 0$  to  $K$  do // For each layer  $i$  in the hierarchy
   $\epsilon_i \leftarrow N(0, I)$ 
   $\hat{\epsilon}_i \leftarrow \frac{\epsilon_i}{\|\epsilon_i\|} \cdot \sqrt{N-1}$ ; // project onto spherical shell
   $\theta_i \leftarrow \text{ToPolarCoords}(\hat{\epsilon}_i)$ 
end
return  $\{\theta_i\}_1^N$ 

```

Algorithm 2: Optimization in Latent Space with Hierarchical Radius Constraint

```

input : Property models:  $[f_1, \dots, f_m]$ , Prior distribution:
           $[p(z_K | NN_K(z_{<K})), \dots, p(z_1 | NN_0(z_0)), p(z_0)]$ , Objective function:  $O(\cdot)$ 
output: Spherical coordinates of a molecule in latent space with converged property values
initialize  $\{\theta_i\}_1^K \leftarrow \text{InitializeAngles}()$ ;
// The first layer of the prior is a standard Gaussian
 $\mu_0 \leftarrow 0, \sigma_0 \leftarrow 1$ ;
for  $i \leftarrow 0$  to  $K$  do // For each layer  $i$  in the hierarchy
   $z_i \leftarrow \text{ToCartesianCoords}(\theta_i)$ 
   $\hat{z}_i \leftarrow z_i \cdot \sigma_i + \mu_i$ ; // Re-parametrize standard Gaussian variable
  to conditional Gaussian at position  $i$  in the hierarchy
   $\cdot \mu_{i+1}, \sigma_{i+1} \leftarrow \text{NN}_i(\hat{z}_{<i})$ ; // Compute  $\mu$  and  $\sigma$  of the next level
end
// Optimize  $\{\theta_i\}$  until the objective function  $O(\cdot)$  has converged
 $\{\theta_i^*\}_1^N \leftarrow \text{GradientDescent}(O(\{f_j\}_1^M, \{z_i(\theta_i)\}_1^K))$ ;
return  $\{\theta_i^*\}_1^K$ 

```

Molecular optimization is quite robust to hyperparameters. We considered ADAM learning rates in $\{0.1, 0.01, 0.001, 0.0001\}$ and $\beta \in \{0.1, 0.01, 0.001, 0.0001\}$.

4.5 Datasets

SMILES strings, as well as the true values of the log octanol-water partition coefficient (logP), molecular weight (MW), and the quantitative estimate of drug-likeness (QED), are computed using RDKit [59].

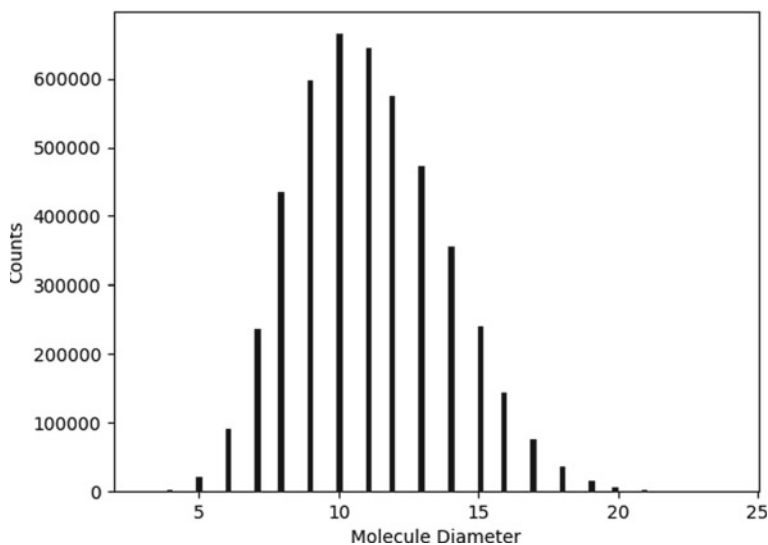


Fig. 4.8 Histogram of molecular diameters in the ZINC250k dataset. The diameter is defined as the maximum eccentricity over all atoms in the molecular graph. The mean is 11.1; the maximum is 24. Typical implementations of graph convolution use only three to seven rounds of message passing [11, 13, 14, 34, 39, 40, 51] and so cannot propagate information across most molecules in this dataset

4.5.1 ZINC

For molecular property optimization and fully supervised property prediction, we train the All SMILES VAE on the ZINC250k dataset of 250,000 organic molecules with between 6 and 38 heavy atoms and penalized logPs from -13 to 5 [8]. This dataset is curated from a subset of the ZINC12 dataset [64] and available from https://github.com/aspuru-guzik-group/chemical_vae. The distribution of molecular diameters in ZINC250k is shown in Fig. 4.8. Penalized logP is commonly used in molecular optimization benchmarks and comprises the log octanol-water partition coefficient minus the synthetic accessibility score and the number of rings with more than six atoms, with all component terms normalized to have zero mean and unit standard deviation on the ZINC250k dataset [9–11, 13, 34, 35].

For semi-supervised property prediction on logP, MW, and QED, we train on the ZINC310k dataset of 310,000 organic molecules with between 6 and 38 heavy atoms [16]. This dataset is curated from the full ZINC15 dataset [65] and available from <https://github.com/nyu-dl/conditional-molecular-design-ssvae>.

4.5.2 *Tox21*

For the semi-supervised prediction of twelve forms of toxicity, we train on the Tox21 dataset [61, 62], accessed through the DeepChem package [66], with the provided random train/validation/test set split. This dataset contains binarized binding affinities against up to 12 proteins for 6264 training, 783 validation, and 784 test molecules. Tox21 contains molecules with up to 140 atoms, ranging from large peptides to lanthanide, actinide, and other metals. Many of these metal atoms are not present in any of the standard molecular generative modeling datasets, and there are metal atoms in the validation and test set that never appear in the training set. To address these difficulties, we curated an unsupervised dataset of 1.5 million molecules from the PubChem database [67]. To maintain commensurability with prior work, this additional unsupervised dataset is *only* used on the Tox21 prediction task.

4.6 Results

We compare the performance of the All SMILES VAE to a variety of state-of-the-art algorithms that have been evaluated on standard molecular property prediction and optimization tasks. In particular, we compare to previously published results on the character/chemical VAE (CVAE) [8] (with results reported in [10]), grammar VAE (GVAE) [10], syntax-directed VAE (SD-VAE) [9], junction tree VAE (JT-VAE) [13], NeVAE [11], semisupervised VAE (SSVAE) [16], graph convolutional policy network (GCPN) [34], molecule deep Q-network (MolDQN) [35], and the DeepChem [66] implementation of extended connectivity fingerprints (ECFP) [68] and graph convolutions (GraphConv) [39, 40, 66]. Extended connectivity fingerprints are a fixed-length hash of local fragments of the molecule, used as input to conventional machine learning techniques such as random forests, support vector machines, and non-convolutional neural networks [66]. For toxicity prediction, we also compare to PotentialNet [69], ToxicBlend [70], and the results of [71].

4.6.1 *Reconstruction Accuracy and Validity*

The full power of continuous, gradient-based optimization can be brought to bear on molecular properties given a bijection between molecules and contractible regions of a latent space, along with a regressor from the latent space to the property of interest that is differentiable almost everywhere. Such a bijection is challenging to confirm, since it is difficult to find the full latent space preimage of a molecule implicitly defined by a mapping from latent space to SMILES strings, such as our beam search decoder. As a necessary condition, we confirm that it is possible to map from the space of molecules to latent space and back again, and that random samples

from the prior distribution in the latent space map to valid molecules. The former is required for injectivity, and the latter for surjectivity, of the mapping from molecules to contractible regions of the latent space.

Using the approximating posterior as the encoder, but always selecting the mean of each conditional Gaussian distribution, and a using beam search over the conditional likelihood as the decoder, $87.4\% \pm 1\%$ of a held-out test set of ZINC250k (80/10/10 train/val/test split) is reconstructed accurately. With the same beam search decoder, $98.5\% \pm 0.1\%$ of samples from the prior decode to valid SMILES strings. We expect that enforcing grammatical constraints in the decoder LSTM, as described in Sect. 4.7, would further increase these rates. All molecules decoded from a set of 50,000 independent samples from the prior were unique, 99.958% were novel relative to the training dataset, and their average synthetic accessibility score [72] was 2.97 ± 0.01 , compared to 3.05 in the ZINC250k dataset used for training.

Previous molecular variational autoencoders have been evaluated using the percentage of molecules that are correctly reconstructed when sampling from both the approximating posterior $q(z|x)$ and the conditional likelihood $p(x|z)$ (reconstruction accuracy), and the percentage of samples from the prior $p(z)$ and conditional likelihood $p(x|z)$ that are valid SMILES strings (validity). While these measures have intuitive appeal, they reflect neither the explicit training objective (the ELBO), nor the requirements of molecular optimization. In particular, when optimizing molecules via the latent space, a deterministic decoder ensures that each point in latent space is associated with a single set of well-defined molecular properties.

The All SMILES VAE is trained on a more difficult task than previous molecular VAEs, since the reconstruction targets are different SMILES encodings than those input to the approximating posterior. This ensures that the latent representation only captures the molecule, rather than its particular SMILES encoding, but it requires the decoder LSTM to produce a complex, highly multimodal distribution over SMILES strings. As a result, samples from the decoder distribution are less likely to correspond to the input to the encoder, either due to syntactic or semantic errors.

To compensate for this unusually difficult decoding task, we evaluate the All SMILES VAE using a beam search over the decoder distribution.⁴ That is, we decode to the single SMILES string estimated to be most probable under the conditional likelihood $p(x|z)$. This has the added benefit of defining an unambiguous decoding for every point in the latent space, simplifying the interpretation of optimization in the latent space (as discussed in Sect. 4.6.3). However, it renders our reconstruction and validity results incommensurable with much prior work, which use stochastic encoders and decoders.

⁴ The full decoder distribution is still used for training.

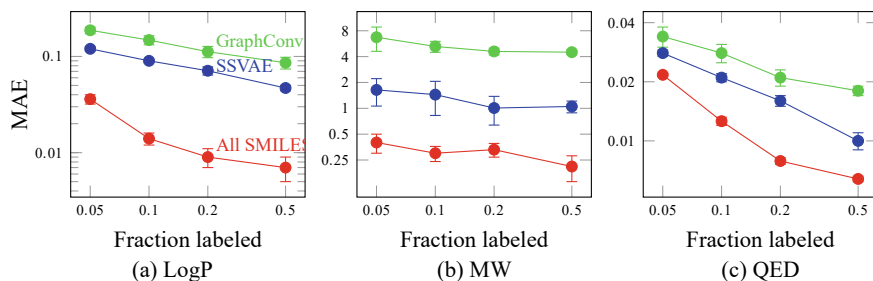


Fig. 4.9 Semi-supervised mean absolute error (MAE) \pm the standard deviation across ten replicates for the log octanol-water partition coefficient (a), molecular weight (b), and the quantitative estimate drug-likeness (c) [60] on the ZINC310k dataset. Plots are log-log; the All SMILES MAE is a fraction of that of the SSSVAE [16] and graph convolutions [40]. Semi-supervised VAE (SSVAE) and graph convolution results are those reported by Kang and Cho [16]

4.6.2 Property Prediction

Ultimately, we would like to optimize molecules for complicated physical properties, such as binding affinity to selected receptors and low toxicity. Networks can only be trained to predict such physical properties if their true values are known on an appropriate training dataset. While simple properties can be accurately computed from first principles, properties like drug efficacy arise from highly nonlinear, poorly characterized processes, and can only be accurately determined through time-consuming and expensive experimental measurements. Since such experiments can only be performed on a small number of molecules, we evaluate the ability of the All SMILES VAE to perform semi-supervised property prediction.

As Fig. 4.9 and Table 4.1 demonstrate, we significantly improve the state of the art in the semi-supervised prediction of simple molecular properties, including the log octanol-water partition coefficient (logP), molecular weight (MW), and quantitative estimate of drug-likeness (QED) [60], against which many algorithms have been benchmarked. We achieve a similar improvement in fully supervised property prediction, as given in Table 4.2, where we compare to extended connectivity fingerprints (ECFP) [68], the character VAE (CVAE) [8], and graph convolutions [39]. We also surpass the state of the art in toxicity prediction on the Tox21 dataset [61, 62], as given in Table 4.2, despite refraining from ensembling our model, or engineering features using expert chemistry knowledge, as in previous state-of-the-art methods [70].

Rather than jointly modeling the space of molecules and their properties, some earlier molecular variational autoencoders first trained an unsupervised VAE on molecules, extracted their latent representations, and then trained a sparse Gaussian process over molecular properties as a function of these fixed latent representations [9–11, 13]. Sparse Gaussian processes are parametric regressors, with the location and value of the inducing points trained based upon the entire supervised dataset [73]. They have significantly more parameters, and are correspondingly more powerful, than linear regressors.

Table 4.1 Mean absolute error (MAE) of semi-supervised property prediction on the log octanol-water partition coefficient (logP), molecular weight (MW), and the quantitative estimate of drug-likeness (QED) on ZINC310k dataset

Model	% labeled	MAE logP	MAE MW	MAE QED
ECFP	50	0.180 ± 0.003	9.012 ± 0.184	0.023 ± 0.000
GraphConv	50	0.086 ± 0.012	4.506 ± 0.279	0.018 ± 0.001
SSVAE	50	0.047 ± 0.003	1.05 ± 0.164	0.01 ± 0.001
All SMILES	50	0.007 ± 0.002	0.21 ± 0.07	0.0064 ± 0.0002
ECFP	20	0.249 ± 0.004	12.047 ± 0.168	0.033 ± 0.001
GraphConv	20	0.112 ± 0.015	4.597 ± 0.419	0.021 ± 0.002
SSVAE	20	0.071 ± 0.007	1.008 ± 0.370	0.016 ± 0.001
All SMILES	20	0.009 ± 0.002	0.33 ± 0.06	0.0079 ± 0.0003
ECFP	10	0.335 ± 0.005	15.057 ± 0.358	0.045 ± 0.001
GraphConv	10	0.148 ± 0.016	5.255 ± 0.767	0.028 ± 0.003
SSVAE	10	0.090 ± 0.004	1.444 ± 0.618	0.021 ± 0.001
All SMILES	10	0.014 ± 0.002	0.30 ± 0.06	0.0126 ± 0.0006
ECFP	5	0.380 ± 0.009	17.713 ± 0.396	0.053 ± 0.001
GraphConv	5	0.187 ± 0.015	6.723 ± 2.116	0.034 ± 0.004
SSVAE	5	0.120 ± 0.006	1.639 ± 0.577	0.028 ± 0.001
All SMILES	5	0.036 ± 0.004	0.4 ± 0.1	0.0217 ± 0.0003

Results other than the All SMILES VAE are those reported by Kang and Cho [16]

Table 4.2 Fully supervised regression on ZINC250k (a), evaluated using the mean absolute error; and Tox21 (b), evaluated with the area under the receiver operating characteristic curve (AUC-ROC), averaged over all 12 toxicity types

(a) ZINC250k			(b) Tox21	
Model	MAE logP	MAE QED	Model	AUC-ROC
ECFP	0.38	0.045	GraphConv + SN	0.854
CVAE	0.15	0.054	PotentialNet	0.857 ± 0.006
CVAE enc	0.13	0.037	ToxicBlend	0.862
GraphConv	0.05	0.017	All SMILES (no harmonization)	0.864 ± 0.003
All SMILES	0.005 ± 0.0006	0.0052 ± 0.0001	All SMILES	0.8751 ± 0.0008

Aside from All SMILES, results are those reported by ECFP: [68], CVAE: [8], GraphConv: [39], Graph Conv + Super Node (SN): [71], PotentialNet: [69], and ToxicBlend: [70]. The ablation of atom harmonization is also evaluated on the Tox21 dataset

Molecular properties are only a smooth function of the VAE latent space when the property regressor is trained jointly with the generative model [8]. Results using a sparse Gaussian process on the latent space of an unsupervised VAE are very poor compared to less powerful regressors trained jointly with the VAE. Our property prediction is two orders of magnitude more accurate than sparse Gaussian process regression on an unsupervised VAE latent representation, as given in Table 4.3.

Table 4.3 Root mean square error of the log octanol-water partition coefficient (logP) on the ZINC250k dataset

Model	RMSE
Character VAE (CVAE) [8, 10]	1.504
Grammar VAE (GVAE) [10]	1.404
Syntax-directed VAE (SD-VAE) [9]	1.366
Junction tree VAE (JT-VAE) [13]	1.290
NeVAE [11]	1.23
All SMILES	0.011 ± 0.001

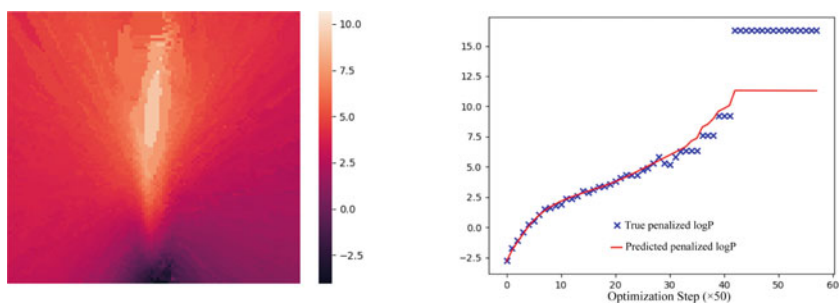
Results other than the All SMILES VAE are those reported in the cited papers

Accurate property prediction only facilitates effective optimization if the true property value is smooth with respect to the latent space. In Fig. 4.10a, we plot the true (not predicted) logP over a densely sampled 2D slice of the latent space, where the y-axis is aligned with the logP linear regressor.

Pathways on which activity (active or inactive) is assessed for the Tox21 dataset include seven nuclear receptor signaling pathways: androgen receptor, full (NR-AR); androgen receptor, LBD (NR-AR-LBD); aryl hydrocarbon receptor (NR-AHR); aromatase (NR-AROMATASE); estrogen receptor alpha, LBD (NR-ER-LBD); estrogen receptor alpha, full (NR-ER); and peroxisome proliferator-activated receptor gamma (NR-PPAR-GAMMA). The Tox21 dataset also includes activity assessments for five stress response pathways: nuclear factor (erythroid-derived 2)-like 2/antioxidant responsive element (SR-ARE); ATAD5 (SR-ATAD5); heat shock factor response element (SR-HSE); mitochondrial membrane potential (SR-MMP); and p53 (SR-p53). We report the area under the receiver operating characteristic curve (AUC-ROC) on each assay independently in Table 4.4. The average of these AUC-ROCs is reported in Table 4.2. We do not include the result of [40] in Table 4.2, since it is not evaluated on the same train/validation/test split of the Tox21 dataset, and so is not commensurable.

4.6.3 Molecular Optimization

We maximize the output of our linear and logistic property regressors, plus a log-prior regularizer, with respect to the latent space, subject to the hierarchical radius constraint described in Sect. 4.4.2. After optimizing in the latent space with ADAM, we project back to a SMILES representation of a molecule with the decoder. Following prior work, we optimize QED and logP penalized by the synthetic accessibility score and the number of large rings [9–11, 13, 34, 35]. Figure 4.10b depicts the predicted and true penalized logP over an optimization trajectory, while Table 4.5 compares the top three values found among 100 such trajectories to the previous state



(a) True logP over a 2D slice of latent space

(b) Predicted and true logP over optimization



(c) Coarse sampling of decoded molecules from a 2D slice of latent space

Fig. 4.10 Dense decodings of true penalized logP along a local 2D sheet in latent space, with the y-axis aligned with the regressor (a), and predicted and true penalized logP across steps of optimization (b). We also display a coarse sampling of the molecules corresponding to the logP heatmap (c)

Table 4.4 Area under the receiver operating characteristic curve (AUC-ROC) per assay on the Tox21 dataset

NR-AR	NR-AR-LBD	NR-AHR	NR-AROMATASE	NR-ER	NR-ER-LBD
0.864	0.921	0.909	0.908	0.719	0.811
NR-PPAR-GAMMA	SR-ARE	SR-ATAD5	SR-HSE	SR-MMP	SR-p53
0.935	0.860	0.870	0.901	0.927	0.882

Table 4.5 Properties of the top three optimized molecules trained on ZINC250k

Model	Penalized logP	Model	QED
JT-VAE	5.30, 4.93, 4.49	JT-VAE	0.925, 0.911, 0.910
GCPN	7.98, 7.85, 7.80	CGVAE	0.938, 0.931, 0.880
MolDQN	8.93, 8.93, 8.91	GCPN	0.948, 0.947, 0.946
All SMILES	12.31, 12.13, 12.01	MolDQN	0.948, 0.948, 0.948
All SMILES (KL unscaled)	29.80, 29.76, 29.11	All SMILES	0.948, 0.948, 0.948

Other results are taken from JT-VAE: [13], GCPN: [34], MolDQN: [35], and CGVAE: [14]. Following prior work, penalized logP is normalized by the statistics of the Zinc250k dataset

of the art.⁵ The molecules realizing these property values are shown in Fig. 4.11. The molecules optimized for penalized logP in Fig. 4.11a are more akin to polymers than small molecules, despite the training set consisting of small molecules from ZINC, reflecting the ability of the model to generalize beyond its training set. We present an optimization trajectory for the quantitative estimate of drug-likeness (QED) in Fig. 4.12.

For the molecules depicted in Fig. 4.11, we scaled $\text{KL}(q(z|x)||p(z))$ in the ELBO (Eq. 4.1) of the All SMILES VAE by the number of SMILES strings in the decoder. This renders the loss function analogous to that of many parallel single-SMILES VAEs, but with message passing between encoders leading to a shared latent representation. If we leave the KL term unscaled, latent space embeddings are subject to less regularization forcing them to match the prior distribution. Optimization of molecular properties with respect to the latent space therefore searches over a wider space of molecules, which are less similar to the training set.

⁵ Zhou et al. [35] appear to report unnormalized penalized logP values: 11.84, 11.84, and 11.82. In Table 4.5, we recompute normalized values for their best molecules. Recently, Winter et al. [28] reported molecules with penalized logP as large as 26.1, but train on an enormous, non-standard dataset of 72 million compounds aggregated from the ZINC15 and PubChem databases.

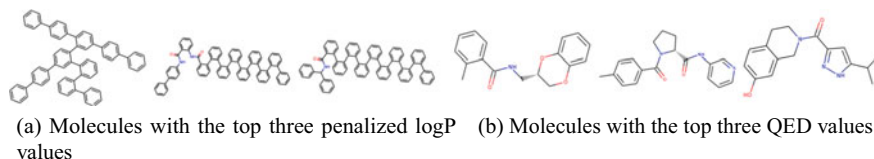


Fig. 4.11 Molecules produced by gradient-based optimization in the All SMILES VAE

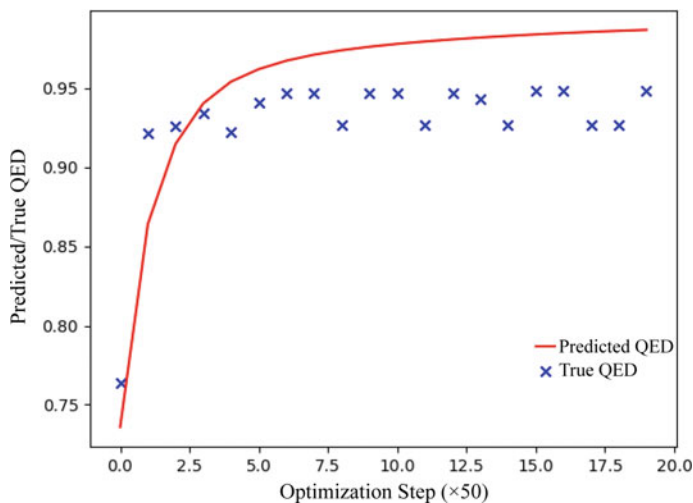


Fig. 4.12 Predicted (red line) and true (blue x's) quantitative estimate of drug-likeness (QED) over the optimization trajectory resulting in the molecule with the maximum observed true QED (0.948)

4.6.4 Ablation of Model Components

In Table 4.6, we progressively ablate model components to demonstrate that all elements of the All SMILES architecture contribute to building a powerful fixed-length representation of molecules, rather than their particular SMILES string instantiations. We evaluate the effect of these ablations on the mean absolute error (MAE) of logP and QED predictions, as well as the percentage of samples from the prior that decode to valid SMILES strings (Val) and the percentage of test molecules that are reconstructed accurately (Rec acc). In all cases, we use the mean of each conditional Gaussian distribution and a beam search decoder.

NO ATOM HARMONIZATION removes the pooling among each instance of an atom across SMILES strings in the encoder, depicted in Fig. 4.4. As a result, the multiple SMILES inputs are processed independently until the final max pooling over GRU hidden states. A random SMILES string is chosen to serve as input to the attention mechanisms of the approximating posterior. Table 4.2b shows the significant effect of this ablation on toxicity prediction, demonstrating the importance of atom harmo-

Table 4.6 Effect of model ablation on fully supervised property prediction and generative modeling using the ZINC250k dataset

Ablation	MAE logP	MAE QED	Val	Rec acc
Full model	0.005 \pm 0.0006	0.0052 \pm 0.0001	98.5 \pm 0.1	87.4 \pm 1.0
No atom harmonization	0.008 \pm 0.004	0.0076 \pm 0.0005	97.6 \pm 0.2	84.0 \pm 0.4
One SMILES enc	0.008 \pm 0.005	0.0073 \pm 0.0002	98.4 \pm 0.1	82.3 \pm 0.4
One SMILES enc/dec (\neq)	0.009 \pm 0.001	0.0091 \pm 0.0003	97.1 \pm 0.7	80.9 \pm 0.4
One SMILES enc/dec (=)	0.025 \pm 0.003	0.0115 \pm 0.0004	85.7 \pm 1	91.3 \pm 0.6
No posterior hierarchy	0.010 \pm 0.003	0.0051 \pm 0.0001	98.2 \pm 0.5	85.2 \pm 0.6

Table 4.7 Effect of the hierarchical radius constraint on penalized logP optimization

Ablation	First best logP	Second best logP	Third best logP
With radius constraint	17.0 \pm 3.0	16.0 \pm 2.0	14.8 \pm 0.3
Without radius constraint	8.5044 \pm 0.0	6.9526 \pm 0	5.36 \pm 0.05

Predicted penalized logP was evaluated on 1000 optimization trajectories. From these, the true logP was evaluated on the 100 best trajectories, and the top three true penalized logPs are reported. Each optimization was repeated 5 times

nization for nonlinear properties of the entire molecule, in contrast to the quasi-linear logP and QED reported in Table 4.6. We extend this process in ONE SMILES ENC by only feeding a single SMILES string to the encoder, although the decoder still reconstructs multiple disjoint SMILES strings. ONE SMILES ENC/DEC (\neq) further reduces the size of the decoder set to one, but the encoded and decoded SMILES strings are distinct. Finally, ONE SMILES ENC/DEC (=) encodes and decodes a single, shared SMILES string. Except for ONE SMILES ENC/DEC (=), all of these ablations primarily disrupt the flow of messages between the flattened spanning trees and induce a similar, significant decay in performance. ONE SMILES ENC/DEC (=) further permits the latent representation to encode the details of the particular SMILES string, rather than forcing the representation of only the underlying molecule, and causes a further reduction in performance.

We also observe a meaningful contribution from the hierarchical approximating posterior. In NO POSTERIOR HIERARCHY, we move all latent variables to the first layer of the hierarchy, removing the succeeding layers. The remaining prior is a standard Gaussian, and there is no attentional pooling over the atom representations.

Table 4.7 shows that the hierarchical radius constraint significantly improves molecular optimization. In contrast to Table 4.5, optimization is performed on penalized logP alone, without a log prior regularizer. This produces better results without the radius constraint and so constitutes a more conservative ablation experiment.

4.7 SMILES Grammar Can Be Enforced with a Pushdown Automaton

The subset of the SMILES grammar [7] captured by Dai et al. [9] and Kusner et al. [10] is equivalent to the context-free grammar as shown in Fig. 4.13. This subset does not include the ability to represent multiple disconnected molecules in a single SMILES string, multiple fragments that are only connected by ringbonds, or wildcard atoms. `element_symbols` includes symbols for every element in the periodic table, including the `aliphatic_organic` symbols.

Productions generally begin with a unique, defining symbol or set of symbols. Exceptions include `bond` and `charge` (both can begin with `-`), and `aliphatic_organic` and `aromatic_symbols` (both include `c`, `n`, `o`, `s`, and `p`), but these pairs of productions never occur in the same context, and so cannot be confused. The particular production for `chiral` can only be resolved by parsing characters up to the next production, but the end of `chiral` and the identity of the subsequent production can be inferred from its first symbol of the production after `chiral`. Alternatively, the strings of `chiral` can be encoded as monolithic tokens.

Whenever there is a choice between productions, the true production is uniquely identified by the next symbols. The only aspect of the SMILES grammar that requires

```

chain → branched_atom rest_of_chain
rest_of_chain → ε | bond? chain
bond → '-' | '=' | '#' | '$' | ':' | '/' | '\'
branched_atom → atom ringbond* branch*
ringbond → bond digit? digit
branch → '(' bond? chain ')'
atom → aliphatic_organic | aromatic_organic | bracket_atom
aliphatic_organic → 'B' | 'C' | 'N' | 'O' | 'S' | 'P' | 'F' | 'Cl' | 'Br' | 'I'
aromatic_organic → 'b' | 'c' | 'n' | 'o' | 's' | 'p'
bracket_atom → '[' isotope? symbol chiral? hcount? charge? class? ']'
isotope → digit? digit? digit
symbol → element_symbols | aromatic_symbols
aromatic_symbols → 'c' | 'n' | 'o' | 'p' | 's' | 'se' | 'as'
chiral → '@' | '@@' | '@TH1' | '@TH2' | '@AL1' | '@AL2' |
        '@SP1' | '@SP2' | '@SP3' | '@TB1' | '@TB2' ... '@TB30' |
        '@OH1' | '@OH2' ... '@OH30'
hcount → 'H' digit?
charge → '-' digit? | '+' digit?
class → ':' digit? digit? digit?
digit → '0' | '1' | '2' | '3' | '4' | '5' | '6' | '7' | '8' | '9'

```

Fig. 4.13 Context-free grammar of SMILES strings

more than a few bits of memory is the matching of parentheses, which can be performed in a straightforward manner with a pushdown automaton. As a result, parse trees [9, 10] need not be explicitly constructed by the decoder to enforce the syntactic restrictions of SMILES strings. Rather, the SMILES grammar can be enforced with a pushdown automaton running in parallel with the decoder RNN. The state of the pushdown automaton tracks progress within the representation of each atom and the sequence of atoms and bonds. The output symbols available to the decoder RNN are restricted to those consistent with the current state of the pushdown automaton. (and [are pushed onto the stack when emitted and must be popped from the top of the stack in order to emit) or] respectively.

For example, in addition to simple aliphatic organic (B, C, N, O, S, P, F, Cl, Br, or I) or aromatic organic (b, c, n, o, s, or p) symbols, an atom may be represented by a pair of brackets (requiring parentheses matching) containing a sequence of isotope number, atom symbol, chiral symbol, hydrogen count, charge, and class. With the exception of the atom symbol, each element of the sequence is optional, but is easily parsed by a finite state machine. `isotope`, `symbol`, `chiral`, `hcount`, `charge`, and `class` can all be distinguished based upon their first character, so the position in the progression can be inferred trivially.⁶

When parsing `branched_atom`, all productions after the initial `atom` are `ringbonds` until the first (, which indicates the beginning of a branch. After observing a), and popping the complementary (off of the stack, the SMILES string is necessarily in the third component of a `branched_atom`, since only a `branched_atom` can emit a branch, and only `branch` produces the symbol). The next symbol must be a (, indicating the beginning of another branch, or one of the first symbols of `rest_of_chain`, since this must follow the `branched_atom` in the `chain` production.

4.7.1 Ringbond and Valence Shell Semantic Constraints

Similarly, the semantic restrictions of ringbond matching and valence shell constraints can be enforced during feedforward production of a SMILES string using a pushdown stack and a small (100-element) random access memory. Our approach depends upon the presence of matching bond labels at both sides of a ringbond, which is allowed but not required in standard SMILES syntax. We assume the trivial extension of the SMILES grammar to include this property.

`ringbonds` are constrained to come in pairs, with the same bond label on both sides. Whenever a given `ringbond` is observed, flip a bit in the random access memory corresponding to the ring number (the set of digits after the `bond`). When the `ringbond` bit is flipped on, record the associated `bond` in the random access memory associated with the ring number; when the `ringbond` bit is flipped off, require that the new `bond` matches the recorded `bond`, and clear the random access

⁶ `symbol` and `hcount` can both start with 'H', but `symbol` is mandatory, so there is no ambiguity.

memory of the bond. The molecule is only allowed to terminate (`rest_of_chain` produces ϵ rather than `bond? chain`) when all `ringbond` bits are off (parity is even). The decoder may receive as input which `ringbonds` are open and the associated `bond` type, so it can preferentially close them.

The set of nested atomic contexts induced by `chain`, `branched_atom`, and `branch` can be arbitrarily deep, corresponding to the depth of branching in the spanning tree realized by a SMILES string. As a result, the set of SMILES symbols describing bonds to a single atom can be arbitrarily far away from the associated atom. However, once a branch is entered, it must be traversed in its entirety before the SMILES string can return to the parent atom. For each atom, it is sufficient to push the valence shell information onto the stack as it is encountered. If the SMILES string enters a branch while processing an atom, simply push on a new context, with a new associated root atom. Once the branch is completed, pop this context off the stack and return to the original atom.

More specifically, each atom in the molecule is completely described by a single `branched_atom` and the bond preceding it (from the `rest_of_chain` that produced the `branched_atom`). Within each successive pair of `bond` and `branched_atom`, track the sum of the incoming `rest_of_chain` bond, the internal `ringbond` and `branch` bonds, and outgoing `rest_of_chain` bond (from the succeeding `rest_of_chain`) on the stack. That is, each time a new bond is observed from the atom, pop off the old valence shell count and push on the updated count. Require that the total be less than a bound set by the atom, any remaining bonds are filled by implicit hydrogen atoms. Provide the number of available bonds as input to the decoder RNN, and mask additional `ringbonds` and `branches` once the number of remaining available bonds reaches one (if there are still open `ringbonds`) or zero (if all `ringbonds` are closed). Mask the outgoing bond, or require that `rest_of_chain` produce ϵ , based upon the number of remaining available bonds.

4.7.2 *Redundancy in Graph-Based and SMILES Representations of Molecules*

To avoid the degeneracy of SMILES strings, for which there are many encodings of each molecule, some authors have advocated the use of graph-based representations [14, 21–23]. While graph-based processing may produce a unique representation in the encoder, it is not possible to avoid degeneracy in the decoder. Parse trees [9, 10], junction trees [13], lists of nodes and edges [11, 14, 22], and vectors/matrices of node/edge labels [20, 21, 23] all imply an ordering among the nodes and edges, with many orderings describing the same graph. Canonical orderings can be defined, but unless they are obvious to the decoder, they make generative modeling harder rather than easier, since the decoder must learn the canonical ordering rules. Graph matching procedures can ensure that probability within a generative model is assigned to

the correct molecule, regardless of the order produced by the decoder [23]. However, they do not eliminate the degeneracy in the decoder's output, and the generative loss function remains highly multimodal.

4.8 Conclusion

For each molecule, the All SMILES encoder uses stacked, pooled RNNs on multiple SMILES strings to efficiently pass information throughout the molecular graph. The decoder targets a disjoint set of SMILES strings of the same molecule, forcing the latent space to develop a consistent representation for each molecule. Attentional mechanisms in the approximating posterior summarize spatially diffuse features into a fixed-length, non-factorial approximating posterior, and construct a latent representation on which linear regressors achieve state-of-the-art semi- and fully supervised property prediction. Gradient-based optimization of these regressor outputs with respect to the latent representation, constrained to a subspace near almost all probability in the prior, produces state-of-the-art optimized molecules when coupled with a simple RNN decoder.

References

1. Pyzer-Knapp EO, Suh C, Gómez-Bombarelli R, Aguilera-Iparraguirre J, Aspuru-Guzik A (2015) What is high-throughput virtual screening? A perspective from organic materials discovery. *Ann Rev Mater Res* 45:195–216
2. Bohacek RS, McMartin C, Guida WC (1996) The art and practice of structure-based drug design: a molecular modeling perspective. *Med Res Rev* 16(1):3–50
3. Stumpfe D, Bajorath J (2012) Exploring activity cliffs in medicinal chemistry: miniperspective. *J Med Chem* 55(7):2932–2942
4. Sanchez-Lengeling B, Aspuru-Guzik A (2018) Inverse molecular design using machine learning: generative models for matter engineering. *Science* 361(6400):360–365
5. Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv preprint [arXiv:1312.6114](https://arxiv.org/abs/1312.6114)
6. Rezende DJ, Mohamed S, Wierstra D (2014) Stochastic backpropagation and approximate inference in deep generative models. In: International conference on machine learning, pp 1278–1286
7. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 28(1):31–36
8. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A (2018) Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent Sci* 4(2):268–276
9. Dai H, Tian Y, Dai B, Skiena S, Song L (2018) Syntax-directed variational autoencoder for structured data. arXiv preprint [arXiv:1802.08786](https://arxiv.org/abs/1802.08786)
10. Kusner MJ, Paige B, Hernández-Lobato JM (2017) Grammar variational autoencoder. arXiv preprint [arXiv:1703.01925](https://arxiv.org/abs/1703.01925)
11. Samanta B, De A, Jana G, Chattaraj PK, Ganguly N, Gomez-Rodriguez M (2018) NeVAE: a deep generative model for molecular graphs. arXiv preprint [arXiv:1802.05283](https://arxiv.org/abs/1802.05283)

12. Aumentado-Armstrong T (2018) Latent molecular optimization for targeted therapeutic design. arXiv preprint [arXiv:1809.02032](https://arxiv.org/abs/1809.02032)
13. Jin W, Barzilay R, Jaakkola T (2018) Junction tree variational autoencoder for molecular graph generation. arXiv preprint [arXiv:1802.04364](https://arxiv.org/abs/1802.04364)
14. Liu Q, Allamanis M, Brockschmidt M, Gaunt AL (2018) Constrained graph variational autoencoders for molecule design. arXiv preprint [arXiv:1805.09076](https://arxiv.org/abs/1805.09076)
15. Mueller J, Gifford D, Jaakkola T (2017) Sequence to better sequence: continuous revision of combinatorial structures. In: International conference on machine learning, pp 2536–2544
16. Kang S, Cho K (2018) Conditional molecular design with deep generative models. arXiv preprint [arXiv:1805.00108](https://arxiv.org/abs/1805.00108)
17. Lim J, Ryu S, Kim JW, Kim WY (2018) Molecular generative model based on conditional variational autoencoder for de novo molecular design. *J Cheminformatics* 10(1):31
18. Gupta A, Müller AT, Huisman BJ, Fuchs JA, Schneider P, Schneider G (2018) Generative recurrent networks for de novo drug design. *Mol Inform* 37(1–2):1700111
19. Segler MH, Kogej T, Tyrchan C, Waller MP (2017) Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent Sci* 4(1):120–131
20. De Cao N, Kipf T (2018) Molgan: an implicit generative model for small molecular graphs. arXiv preprint [arXiv:1805.11973](https://arxiv.org/abs/1805.11973)
21. Ma T, Chen J, Xiao C (2018) Constrained generation of semantically valid graphs via regularizing variational autoencoders. In: Advances in neural information processing systems, pp 7113–7124
22. Li Y, Vinyals O, Dyer C, Pascanu R, Battaglia P (2018) Learning deep generative models of graphs. arXiv preprint [arXiv:1803.03324](https://arxiv.org/abs/1803.03324)
23. Simonovsky M, Komodakis N (2018) Graphvae: towards generation of small graphs using variational autoencoders. In: International conference on artificial neural networks. Springer, Berlin, pp 412–422
24. Vinyals O, Bengio S, Kudlur M (2015) Order matters: sequence to sequence for sets. arXiv preprint [arXiv:1511.06391](https://arxiv.org/abs/1511.06391)
25. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Advances in neural information processing systems, pp 3104–3112
26. Bjerrum EJ (2017) Smiles enumeration as data augmentation for neural network modeling of molecules. arXiv preprint [arXiv:1703.07076](https://arxiv.org/abs/1703.07076)
27. Bjerrum E, Sattarov B (2018) Improving chemical autoencoder latent space and molecular de novo generation diversity with heteroencoders. *Biomolecules* 8(4):131
28. Winter R, Montanari F, Steffen A, Briem H, Noé F, Clevert DA (2019) Efficient multi-objective molecular optimization in a continuous latent space. *Chem Sci* 10(34):8016–8024
29. Winter R, Montanari F, Noé F, Clevert DA (2019) Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem Sci* 10(6):1692–1701
30. Guimaraes GL, Sanchez-Lengeling B, Outeiral C, Farias PLC, Aspuru-Guzik A (2017) Objective-reinforced generative adversarial networks (organ) for sequence generation models. arXiv preprint [arXiv:1705.10843](https://arxiv.org/abs/1705.10843)
31. Jaques N, Gu S, Bahdanau D, Hernández-Lobato JM, Turner RE, Eck D (2017) Sequence tutor: conservative fine-tuning of sequence generation models with kl-control. In: Proceedings of the 34th international conference on machine learning, pp 1645–1654
32. Olivecrona M, Blaschke T, Engkvist O, Chen H (2017) Molecular de-novo design through deep reinforcement learning. *J Cheminformatics* 9(1):48
33. Putin E, Asadulaev A, Ivanenkov Y, Aladinskiy V, Sanchez-Lengeling B, Aspuru-Guzik A, Zhavoronkov A (2018) Reinforced adversarial neural computer for de novo molecular design. *J Chem Inf Model* 58(6):1194–1204
34. You J, Liu B, Ying R, Pande V, Leskovec J (2018) Graph convolutional policy network for goal-directed molecular graph generation. arXiv preprint [arXiv:1806.02473](https://arxiv.org/abs/1806.02473)
35. Zhou Z, Kearnes S, Li L, Zare RN, Riley P (2018) Optimization of molecules via deep reinforcement learning. arXiv preprint [arXiv:1810.08678](https://arxiv.org/abs/1810.08678)

36. Popova M, Isayev O, Tropsha A (2018) Deep reinforcement learning for de novo drug design. *Sci Adv* 4(7):eaap7885
37. Kipf TN, Welling M (2016) Variational graph auto-encoders. arXiv preprint [arXiv:1611.07308](https://arxiv.org/abs/1611.07308)
38. Lusci A, Pollastri G, Baldi P (2013) Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *J Chem Inf Model* 53(7):1563–1575
39. Duvenaud DK, Maclaurin D, Iparraguirre J, Bombarell R, Hirzel T, Aspuru-Guzik A, Adams RP (2015) Convolutional networks on graphs for learning molecular fingerprints. In: *Advances in neural information processing systems*, pp 2224–2232
40. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P (2016) Molecular graph convolutions: moving beyond fingerprints. *J Comput Aided Mol Des* 30(8):595–608
41. Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv preprint [arXiv:1609.02907](https://arxiv.org/abs/1609.02907)
42. Li Y, Tarlow D, Brockschmidt M, Zemel R (2015) Gated graph sequence neural networks. arXiv preprint [arXiv:1511.05493](https://arxiv.org/abs/1511.05493)
43. Ryu S, Lim J, Kim WY (2018) Deeply learning molecular structure-property relationships using graph attention neural network. arXiv preprint [arXiv:1805.10988](https://arxiv.org/abs/1805.10988)
44. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9(8):1735–1780
45. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint [arXiv:1406.1078](https://arxiv.org/abs/1406.1078)
46. Hammond DK, Vandergheynst P, Gribonval R (2011) Wavelets on graphs via spectral graph theory. *Appl Comput Harmonic Anal* 30(2):129–150
47. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*, pp 1097–1105
48. LeCun Y, Boser BE, Denker JS, Henderson D, Howard RE, Hubbard WE, Jackel LD (1990) Handwritten digit recognition with a back-propagation network. In: *Advances in neural information processing systems*, pp 396–404
49. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 770–778
50. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 2818–2826
51. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE (2017) Neural message passing for quantum chemistry. In: *Proceedings of the 34th international conference on machine learning*, vol 70, pp 1263–1272
52. Clayden J, Greeves N, Warren S, Wothers P (2001) *Organic chemistry*. Oxford University Press, Oxford
53. Im Y, Kim M, Cho YJ, Seo JA, Yook KS, Lee JY (2017) Molecular design strategy of organic thermally activated delayed fluorescence emitters. *Chem Mater* 29(5):1946–1963
54. Ba JL, Kiros JR, Hinton GE (2016) Layer normalization. arXiv preprint [arXiv:1607.06450](https://arxiv.org/abs/1607.06450)
55. Dauphin YN, Fan A, Auli M, Grangier D (2017) Language modeling with gated convolutional networks. In: *Proceedings of the 34th international conference on machine learning*, vol 70, pp 933–941
56. Rolfe JT (2016) Discrete variational autoencoders. arXiv preprint [arXiv:1609.02200](https://arxiv.org/abs/1609.02200)
57. Ioffe S (2017) Batch renormalization: towards reducing minibatch dependence in batch-normalized models. In: *Advances in neural information processing systems*, pp 1945–1953
58. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint [arXiv:1409.0473](https://arxiv.org/abs/1409.0473)
59. Landrum G et al (2006) Rdkit: open-source cheminformatics
60. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL (2012) Quantifying the chemical beauty of drugs. *Nat Chem* 4(2):90

61. Huang R, Xia M, Nguyen DT, Zhao T, Sakamuru S, Zhao J, Shahane SA, Rossoshek A, Simeonov A (2016) Tox21challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Front Environ Sci* 3:85
62. Mayr A, Klambauer G, Unterthiner T, Hochreiter S (2016) Deeptox: toxicity prediction using deep learning. *Front Environ Sci* 3:80
63. Blum A, Hopcroft J, Kannan R (2017) Foundations of data science. <https://www.microsoft.com/en-us/research/publication/foundations-of-data-science-2/>
64. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG (2012) Zinc: a free tool to discover chemistry for biology. *J Chem Inf Model* 52(7):1757–1768
65. Sterling T, Irwin JJ (2015) Zinc 15-ligand discovery for everyone. *J Chem Inf Model* 55(11):2324–2337
66. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V (2018) MoleculeNet: a benchmark for molecular machine learning. *Chem Sci* 9(2):513–530
67. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B et al (2018) Pubchem 2019 update: improved access to chemical data. *Nucleic Acids Res* 47(D1):D1102–D1109
68. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50(5):742–754
69. Feinberg EN, Sur D, Wu Z, Husic BE, Mai H, Li Y, Sun S, Yang J, Ramsundar B, Pande VS (2018) Potentialnet for molecular property prediction. *ACS Central Sci* 4(11):1520–1530
70. Zaslavskiy M, Jégou S, Tramel EW, Wainrib G (2019) Toxicblend: virtual screening of toxic compounds with ensemble predictors. *Comput Toxicol* 10:81–88
71. Li J, Cai D, He X (2017) Learning graph-level representation for drug discovery. arXiv preprint [arXiv:1709.03741](https://arxiv.org/abs/1709.03741)
72. Ertl P, Schuffenhauer A (2009) Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J Cheminformatics* 1(1):8
73. Snelson E, Ghahramani Z (2006) Sparse Gaussian processes using pseudo-inputs. In: *Advances in neural information processing systems*, pp 1257–1264

Chapter 5

SMILES-Based Bioactivity Descriptors to Model the Anti-dengue Virus Activity: A Case Study



Soumya Mitra, Sumit Nandi, Amit Kumar Halder,
and M. Natalia D. S. Cordeiro

Abstract The present work aims to demonstrate the significance of the newly suggested bioactivity descriptors (so-called *signaturizers*) towards developing predictive 2D-QSAR models. As a case study, we examined the development of 2D-QSAR models based on a dataset containing 77 compounds with inhibitory activity reported in a DENV2ProHeLa assay, which is basically a cell-based assay that estimates the Dengivirus-2 (DENV-2) protease inhibitory potential within cellular atmosphere. Indeed, though dengue is a well-known neglected tropical disease, its global incidence has risen sharply in recent years. Moreover, DENV infections may lead to serious and life-threatening diseases such as haemorrhagic fever and dengue shock syndrome. Inhibition of the DENV protease may therefore be a potential target for discovering anti-DENV agents. Interestingly, our initial attempts to set up QSAR models based solely on a number of chemicals descriptors coming from a range of different software packages/programs completely failed, since none of these yielded satisfactory statistical results. Hybrid QSAR models were generated also by combining both chemical and biological descriptors. Noteworthy is that the predictive quality of the 2D-QSAR models significantly improved by resorting instead to solely bioactivity descriptors or those combined with chemical descriptors. The comparison analysis carried out in this work certainly shows that bioactivity descriptors can be useful for setting up predictive models to characterise complex

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-28401-4_5.

S. Mitra · S. Nandi · A. K. Halder
Dr. B. C. Roy College of Pharmacy and Allied Health Sciences, Dr. Meghnad Saha Sarani,
Bidhannagar, Durgapur, West Bengal 713206, India

A. K. Halder · M. N. D. S. Cordeiro (✉)
LAQV@REQUIMTE, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal
e-mail: ncordeir@fc.up.pt

biological activity data, but then of course at the expense of their mechanistic interpretation. Simultaneously, this work provides important guidelines to exploit different linear and non-linear model development strategies in a systematic and consistent manner. What is more, it is based on non-commercial open-access tools, programs and web servers, so that the models can be reproduced, and the proposed models' development strategies be easily and productively followed in the near future.

Keywords Dengue virus · Protease inhibitor · QSAR · Descriptor · Signaturizer

5.1 Introduction

The incidence of dengue fever has increased drastically in recent years owing to high population density, poor environment and health management systems as well as due to increased vector distributions [1, 2]. Being endemic in tropical and subtropical regions, dengue has been categorised as a 'neglected tropical disease', but the rise in international travel to those regions led to an increased number of imported dengue cases in Western countries as well [3]. Around 390 million dengue infections occur per year globally though only 30% cases are clinically recognised [4]. Similar to malaria and filariasis, dengue is a mosquito-borne viral disease that typically causes symptoms such as high fevers, headaches, muscle pains and rash. Dengue virus (DENV) belongs to the Flaviviridae family, which are single-positive-stranded RNA viruses. The DENV is transmitted by four major serotypes, namely: DENV-1, DENV-2, DENV-3 and DENV-4. The DENV-carrying female *Aedes* mosquitoes, including *Aedes albopictus* and *Aedes aegypti* may infect humans [2]. One critical disorder caused by dengue infections is thrombocytopenia, which is normal in both gentle and severe cases [5, 6]. The DENV infections may lead to serious and life-threatening diseases such as haemorrhagic fever and dengue shock syndrome [7]. The WHO declared that the reported deaths increased from 960 to 4032 between the year 2000 and 2015. Despite growing threats of dengue, its treatment still remains symptomatic, focusing mainly on the management of fever, pain and body fluid [8]. Even though a vaccine Dengvaxia has recently been developed, it performs differently in seropositive and seronegative patients, and its application is thus highly restricted [9, 10]. In-depth studies are thus required to develop small molecules as potential therapeutic agents to resist DENV infection.

Meanwhile, the protease inhibitors remained one of the most potential targets for antiviral chemotherapy. The flaviviral protease complex (NS2B-NS3) is responsible for the cleavage of the viral polyprotein into separate functional proteins responsible for the replication of viruses [8, 11]. Inhibition of the DENV protease may therefore be a potential target for discovering anti-DENV agents. Recently, Klein and co-workers of the Heidelberg University have designed and synthesised a series of synthetic small molecules as potential inhibitors of NS2B-NS3 in DENV-2. The authors set up a luciferase-based DENV-2 protease reporter system in HeLa cells (DENV2ProHeLa) that was employed to estimate the activity of the compounds

in a cellular environment [8, 12]. In the current study, we collected 77 of such data-points from two reports, where the DENV2ProHeLa activity was expressed in 50% effective concentration (EC_{50}), with the aim of setting up linear QSAR models that may characterise the structural attributes important for the higher anti-DENV property of these compounds [8, 12]. Overall, we followed a conventional 2D-QSAR modelling approach but have also employed distinct categories of descriptors step-by-step with the goal of generating the best predictive and validated models from this dataset. By doing so, we attempt to find the significance of the bioactivity molecular descriptors recently introduced by Bertoni et al. [13], which the authors so-referred to as ‘signaturizers’. In contrast to chemical descriptors, that mainly rely on the chemical attributes of compounds, signaturizers tend to describe their biological profile in terms of numerical values. Specifically, this work focuses on a case study with anti-DENV protease inhibitors that combine the influence of both chemical and bioactivity descriptors in order to develop validated predictive 2D-QSAR models. However, as it will be described in this chapter, our case study highlights the significance of these newly developed (as well as less exploited) bioactivity descriptors for setting up predictive models.

5.2 Materials and Methods

5.2.1 Importance of Bioactivity Descriptors

The 2D-QSAR modelling primarily relies on chemical descriptors that represent physicochemical and structural properties of small molecules. Due to availability of large bioactivity databases, it is now possible to set up other numerical representations of molecules beyond chemical structures by detecting their biological properties. Bioactivity signatures are multidimensional vectors that capture 25 different biological traits of the molecule (including target profiles, cellular response and clinical outcomes) in a numerical vector format that is similar to the structural descriptors or fingerprints used in the field of cheminformatics [13]. The source of bioactivity signatures is Chemical Checker (CC) [14], which is an integration of major chemogenomics and drug databases containing 25 different elements ranging from A1–E5 (A: chemistry, B: targets, C: networks, D: cells, E: clinics). The details of their sublabels are shown in Table 5.1.

In CC, each molecule is annotated with multiple n -dimensional vectors (i.e., bioactivity signatures) with respect to the spaces for which experimental information is available. Evidently, all these elements do not have the same number of available data and in fact significant differences exist. However, since these bioactivity spaces are correlated, signatures for any novel compound may be obtained by tackling the metric learning problem using the Siamese neural network (SNN) containing a stacked array of CC signatures available for the compound (belonging to any of the A1–E5 layers: S_i) as input whereas a n -dimensional embedding optimised to distinguish between

Table 5.1 Summary of Chemical Checker spaces

Space	Description	Sublabels	Name
A	Chemistry	A1	2D fingerprints
		A2	3D fingerprints
		A3	Scaffolds
		A4	Structural keys
		A5	Physicochemical parameters
B	Targets	B1	Mechanisms of action
		B2	Metabolic genes
		B3	Crystals
		B4	Binding
		B5	High-throughput screening bioassays
C	Network	C1	Small-molecule roles
		C2	Small-molecule pathways
		C3	Signalling pathways
		C4	Biological processes
		C5	Interactome
D	Cells	D1	Gene expression
		D2	Cancer cell lines
		D3	Chemical genetics
		D4	Morphology
		D5	Cell bioassays
E	Clinics	E1	Therapeutic areas
		E2	Indications
		E3	Side effects
		E4	Diseases and toxicology
		E5	Drug–drug interactions

similar and dissimilar molecules in S_i as output [13]. More specifically, the SNN is fed with triplets of molecules (an anchor molecule, one that is similar to the anchor—i.e., positive, and one that is not—i.e., negative), and the SNN is expected to correctly classify this pattern with a distance measurement based on Euclidean distances computed in the embedding space. Therefore, the 25 SNNs are trained on the basis of existing CC signature molecule triplets reflecting S_i similarities. The SNN embedding of 128 is chosen for all CC space to get an output of 128 dimensions and with ‘global’ option 3200 (= 25 × 128) biological signatures are obtained for each molecule. In the present work, we calculated these global signatures for each dataset compound to build 2D-QSAR models.

5.2.2 Dataset Collection

The EC₅₀ (in μM) values of cell-based DENV2ProHeLa activity of 77 compounds were collected from the literature and these were then converted to pEC₅₀ (in M) and subsequently used as response variables. The SMILES notation and the reported biological activity of the dataset compounds is given in the supplementary materials (Table S1). In-depth details of this comparatively novel assay method can be found elsewhere [12]. Briefly, the DENV2ProHeLa assay, also named as DENV-2 protease reporter gene assay, is basically a high-throughput screening (HTS)-capable intracellular DENV-2 protease assay with luciferase reporter system that enables us to estimate the DENV-2 protease activity in a cellular atmosphere. The assay results also reflect membrane permeability, metabolic stability and cytotoxicity of the compounds under investigation. Since the protease in DENV2ProHeLa cells interacts with a number of human host proteins and membranes, this assay provides biologically more meaningful environment as compared to the biochemical assay conducted with isolated protease. The SMILES structures of the 77-dataset compounds were directly collected from the reports of Klein et al. [8, 12], and these were then converted into 3D.sdf formats using the Discovery Studio Visualizer.

5.2.3 Calculation of Molecular Descriptors

The 3D structures of these compounds were submitted to the OCHEM webserver [15] for the calculation of molecular descriptors. This work resorts to a range of different theoretical chemical descriptors other than biological signatures with attempts to generate statistically reliable models. We looked in the OCHEM webserver [15] for a number of well-known software packages to calculate the molecular descriptors for the dataset compounds, including the following ones: (a) AlvaDesc v.2.0.4 [16], (b) CDK 2.7.1 [17], (c) RDKit (<https://www.rdkit.org/docs/>), (d) simplex representation of molecular structure—SIRMS (<https://github.com/DrrDom/sirms>) [18], (e) ISIDA fragments and GSFragment [19], (f) multilevel neighbourhoods of atoms (MNA) [20], (g) Mera + Mersy [21], (h) Mordred descriptors [22], and (i) PyDescriptors [23]. The application of so many diverse types of descriptors basically aimed to check which descriptors are more capable of generating validated and predictive models. In OCHEM, the structures are first pre-processed using Chemaxon following steps such as standardisation, neutralise, remove salts and clean structures [24]. For calculation of 3D structures, optimisation of the compounds geometries was performed using the Corina tool under the OCHEM platform. The ‘global’ signaturizer descriptors were calculated with signaturizer tool (accessed from <https://gitlab.bnb.irbbarcelona.org/packages/signaturizer>), where the SMILES notation of the several structures was submitted as inputs for the calculation of the descriptors using Jupyter notebook provided with this tool [13].

5.2.4 Development of Linear 2D-QSAR Models

Since the present work compares the performance of a number of descriptor calculating tools (e.g., alvaDesc, PyDescriptors) for setting up predictive and validated linear models, we needed robust but consistent model development strategies. We initially used our *in-house* SFS-QSAR tool (accessed from <https://github.com/ncordeirfcup/SFS-QSAR-tool>) for developing multiple models using the sequential forward selection (SFS) technique [25], as illustrated in Fig. 5.1.

Each dataset containing the response variables and the descriptors were randomly divided into three training set-test set combinations, using in the SFS-QSARtool random seed values of 3, 20 and 42. For each division, the following four scoring functions were applied: the determination coefficient (R^2), the negative mean absolute error (NMAE), the negative mean Poisson deviance (NMPD) and the negative mean gamma deviance (NMGD). Similarly, for each of these scoring functions, two cross-validation strategies were used for model development, namely: (i) no cross-validation and (ii) fivefold cross-validation. Therefore, for each descriptor calculating tool, a total of 24 ($= 3 \times 4 \times 2$) models were generated (see Fig. 5.1). After developing these models, the statistical quality of each model was assessed on the basis of internal and external predictivities, as explained later in this chapter. The data division that produced the best statistical result from SFS was then utilised for generating genetic algorithm-based multiple linear regression (GA-MLR) models by employing the GeneticAlgorithm v.4.1_2 [26]. In contrast to SFS, GA is a stochastic feature selection technique and the latter is based on random selection of the set of descriptors, estimation of fitting scores of these random models followed by cross-over and mutation schemes to improve the fitting scores when setting up the final models [26]. The SFS technique, meanwhile, is a non-stochastic technique that includes descriptors in the model one by one following specific scoring functions, and given the same dataset and parameter settings for model development, the users end up with the same model every time [27]. Descriptor pre-treatment was carried

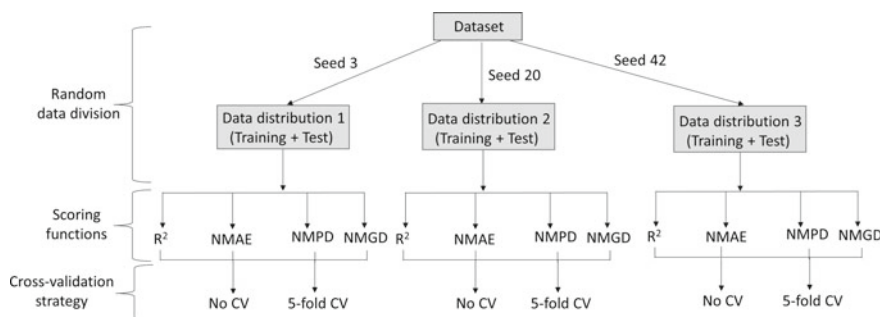


Fig. 5.1 SFS-QSAR model development strategies for each dataset

out for each model development in which constant, near-constant and highly correlated descriptors were eliminated by setting a variance cut-off of 0.0001 and an inter correlation cut-off of 0.99. Notice that even though a high intercorrelation cut-off was employed, the maximum intercollinearity of the final models was carefully checked to ensure that the descriptors of the model are unique and independent. Apart from these, the following parameter settings were used for GA-MLR: (a) total number of iterations/generations: 100, (b) equation length: 8, (c) mutation probability: 0.3, (d) cross-over probability: 1 (default), (e) initial number of equations generated: 100 (default), (f) number of equations selected in each generation: 30. Since GA-MLR models require multiple runs for selecting the best model, in this work, we ran each model 20 times with the training set, and the best model was then chosen based on the overall higher statistical quality [26].

5.2.5 Statistical Analysis of Models

The goodness of fit, robustness and predictivity of the final 2D-QSAR models were estimated using a range of well-known statistical parameters. Initially, the models' internal predictivity was estimated by Q^2_{LOO} and $r_m^2_{\text{LOO}}$, whereas their external predictivity was assessed from R^2_{Pred} and $r_m^2_{\text{test}}$. The final models were more critically examined by checking the R^2 , R^2_{Adj} , the Fisher's statistics (F -test), and the mean absolute error (MAE) values. Furthermore, along with the $r_m^2_{\text{LOO}}$ and $r_m^2_{\text{test}}$ values, their deviations ($\Delta r_m^2_{\text{LOO}}$ and $\Delta r_m^2_{\text{test}}$) were also determined [28–30]. Three additional parameters R^2_{Test} , k , k' and $|r_0^2 - r'^2_0|$, which belong to Golbraikh and Tropsha's acceptable model, criteria were also considered for checking the external predictivity of the test set [30, 31].

As discussed before, the proposed models were checked for intercollinearity, and at the same time, the multicollinearity of the final models was estimated by calculating the variation inflation factor (VIF) using the following equation.

$$\text{VIF} = 1/(1 - R_i^2) \quad (5.1)$$

In this equation, R_i^2 is the determination coefficient (R^2) determined by regressing the i th descriptor on the other descriptors [32].

Additionally, to confirm that the 2D-QSAR model was not developed by chance, the Y -randomisation test was performed to generate the parameter cR_p^2 that measures the difference between original R^2 and average value of randomised R^2 . 1000 randomised models were generated in this work by scrambling the response values [33].

5.2.6 Applicability Domain of the Models

The applicability domain is basically the chemical-biological space within which the prediction of a specific model is deemed reliable. In this work, the Williams plot (leverage vs. standardised residuals) was obtained to identify structural and response outliers in the linear 2D-QSAR models [34, 35].

5.2.7 Non-linear Model Development

The non-linear models were developed using three well-known machine learning tools namely (a) support vector regression (SVR), (b) random forest regression (RFR) and (c) multilayer perception-based regression (MLPR) using our in-house non-linear regression tool (accessed from <https://github.com/ncordeirfcup/Non-linear-Regression-tools>) that employs scikit-learn algorithms to set up non-linear models. In this work, we also performed hyperparameter optimisation for each machine learning technique and the parameters that were tuned during model development are listed in Table 5.2.

Fivefold cross-validated R^2 and R^2_{Pred} were used to estimate the internal and external predictivity of the non-linear models.

Table 5.2 Parameters optimised during the development of non-linear 2D-QSAR models

Technique	Parameters tuning
RFR	Bootstrap: true/false
	Criterion: Gini, entropy
	Maximum depth: 10, 30, 50, 70, 90, 100, 200, none
	Maximum features: auto, sqrt
	Minimum samples leaf: 1, 2, 4
	Minimum samples split: 2, 5, 10
SVR	Number of estimators: 50, 100, 200, 500
	C: 0.1, 1, 10, 100, 1000
	Gamma: 1, 0.1, 0.01, 0.001
	Kernel: RBF, linear
MLPR	Hidden layer sizes: 100
	Activation: identity, logistic, tanh, relu
	Solver: SGD, Adam
	Alpha: 0.0001, 0.001, 0.01, 1
	Learning rate: constant, adaptive, invscaling

5.3 Results and Discussion

As referred to above, descriptors calculated from eight different tools were applied to establish predictive models from the dataset. Finally, the bioactivity descriptors (signaturizers) were employed for setting up models. At first, we resorted to the SFS feature selection technique since this technique is non-stochastic in nature and with a given dataset and parameter settings yields the same model each time. For each descriptor calculating tool, 24 SFS-QSAR models were developed by varying the data distributions (i.e., using as random seed values 3, 20 and 42), scoring functions (i.e., R^2 , NMAE, NMPD and NMGD) and cross-validation techniques (i.e., none and fivefold), as previously shown in Fig. 5.1. To assess the overall quality of these linear 2D-QSAR models, the average values of Q^2_{LOO} and R^2_{Pred} were also calculated (see Table 5.3).

One thing which is clearly seen from the results in Table 5.3 is that, obtaining a predictive linear 2D-QSAR model based on the current dataset is quite challenging. Indeed, models with poor overall predictivity were obtained for descriptors calculated by means of Mera + Mersy, CDK, GSfrag + ISIDA and MNK, whereas moderate predictability was obtained for the models generated with the descriptors coming from PyDescriptors, Mordred, AlvaDes, SIRMS and RDKit. Finally, it is evident from these results that bioactivity descriptors (i.e., signaturizers) led to a linear model, the statistical predictivity of which is considerably higher (around 20%) than the best models generated with other types of descriptors, which clearly underlines the importance of such descriptors in model generation. We hypothesised that better models may be retrieved from PyDescriptors, Mordred, AlvaDes, SIRMS,

Table 5.3 Summary of the statistical results obtained from SFS-QSAR modelling with molecular descriptors calculated with a number of descriptors calculating software/programs

Descriptors	Random seed	Scoring	Fold	Q^2_{LOO}	R^2_{Pred}	$r_m^2_{\text{LOO}}$	$r_m^2_{\text{test}}$	Average ^a
Mera + Mercy	20	R^2	5	0.288	0.203	0.194	0.162	0.246
CDK	20	NMAE	0	0.332	0.398	0.189	0.318	0.365
GSfrag + ISIDA	3	NMAE	5	0.411	0.463	0.274	0.301	0.437
MNK	3	NMAE	5	0.539	0.360	0.401	0.295	0.449
PyDescriptors	20	R^2	0	0.536	0.498	0.403	0.301	0.517
Mordred	3	R^2	5	0.537	0.550	0.401	0.423	0.544
AlvaDesc	42	NMAE	5	0.526	0.576	0.386	0.468	0.551
SIRMS	3	NMAE	0	0.565	0.570	0.439	0.530	0.567
RDKit	3	R^2	0	0.658	0.540	0.546	0.380	0.599
Signaturizers	20	R^2	0	0.717	0.720	0.615	0.660	0.718

^a Average value of Q^2_{LOO} and R^2_{Pred}

Table 5.4 Summary of statistical results obtained from GA-MLR modelling

Descriptors	Random seed	Q^2_{LOO}	R^2_{Pred}	$r_m^2_{\text{LOO}}$	$r_m^2_{\text{test}}$	Average ^a
Signaturizers	20	0.582	0.468	0.448	0.419	0.525
AlvaDesc	42	0.495	0.62	0.356	0.356	0.558
PyDescriptors	20	0.455	0.505	0.308	0.509	0.480
Mordred	3	0.388	0.312	0.243	0.232	0.350
RDKit	3	0.485	0.503	0.343	0.37	0.494
SIRMS	3	0.523	0.483	0.39	0.261	0.503

^a Average value of Q^2_{LOO} and R^2_{Pred}

RDKit and Signaturizers if other feature selection techniques are explored. Therefore, we selected the data distributions of Table 5.3 for each of these descriptors to set up MLR models by means of the stochastic GA feature selection technique and the results are presented in Table 5.4.

As seen, the GA technique failed to improve the quality of the 2D-QSAR MLR models significantly. Indeed, only the model based on the descriptors computed using AlvaDesc reveals a slight quality improvement. Summing up, the SFS-MLR model based on signaturizer descriptors gave us the most predictive linear 2D-QSAR model, judging from the attained Q^2_{LOO} and R^2_{Pred} values (= 0.717 and 0.720, respectively). The statistical quality of this model is significantly better than the models developed with any other tool. Therefore, the next step to be followed is to merge the chemical descriptors with the biological signatures in order to check if more predictive models can be generated or not. For such purpose, we merged the signaturizer descriptors separately with the descriptors calculated by AlvaDesc, RDKit, SIRMS, PyDescriptors, Mordred and MNA. It should be noticed however that descriptors calculated by the remaining tools, such as CDK, Mera + Mercy, were not included since these produced the least predictive models. The same model development strategy was applied for each set of descriptors, i.e., the best model was picked from 24 initially developed SFS-QSAR models by varying the data distributions, the scoring functions, and the cross-validation schemes. The attained results are provided in Table 5.5.

It is now clearly observed that the combination of biological signatures with chemical descriptors improves the overall predictivity of the models as compared to that of the models developed only with chemical descriptors. More importantly, even though the biological signatures provided the most predictive models among descriptors, they do not afford mechanistic interpretations. Yet, hybrid models developed with both the chemical descriptors and biological signatures are able to unveil by some means mechanistic interpretability. From Table 5.5, it is inferred that the most predictive hybrid model is generated with the AlvaDesc descriptors followed by SIRMS and RDKit descriptors. Noticeably, the model produced with signaturizers and Mordred descriptors has a very low $r_m^2_{\text{test}}$ value (= 0.497) indicating that it does not have satisfactory external predictivity. We also attempted to generate these MLR models by using the GA selection but no better model was retrieved. Additionally,

Table 5.5 Summary of the SFS-QSAR models obtained after combining the chemical descriptors calculated with a number of software packages/programs with biological signatures

Descriptors	Random seed	Score	Fold	Q^2_{LOO}	R^2_{Pred}	$r_m^2_{\text{LOO}}$	$r_m^2_{\text{test}}$	Average ^a
Signaturizers + PyDescriptors	20	R^2	0	0.740	0.647	0.644	0.582	0.693
Signaturizers + MNA	42	NMGD	0	0.761	0.654	0.674	0.540	0.707
Signaturizers + RDKit	20	R^2	0	0.742	0.710	0.647	0.626	0.726
Signaturizers + Mordred	42	R^2	0	0.782	0.673	0.698	0.497	0.728
Signaturizers + SIRMS	20	NMGD	5	0.741	0.722	0.644	0.648	0.731
Signaturizers + AlvaDesc	20	NMGD	0	0.760	0.720	0.668	0.631	0.740

^a Average value of Q^2_{LOO} and R^2_{Pred}

when AlvaDesc and SIRMS descriptors were combined with biological signatures, the statistical quality of the resulting models largely deteriorated, having the best obtained model values of 0.691 and 0.623 for Q^2_{LOO} and R^2_{Pred} , respectively. A detailed description of the four most predictive linear 2D-QSAR models obtained in the present work is given in Table 5.6, and the observed versus predicted activity plots for such models are shown in Fig. 5.2.

The statistical significance of these models was also established by the fact that the maximum intercorrelation obtained from these four models are 0.620 (for Model 1), 0.639 (for Model 2) and 0.511 (for both Models 3 and 4). Furthermore, we determined the VIF value for each model descriptor and found that all values were less than five, indicating that multicollinearity does not exist in these models. Moreover, the Y -randomization tests (1000 runs) carried out for each of such models yielded high cR_p^2 values always (> 0.7), which lead us to conclude that they are indeed unique in nature.

Here, it should be also noticed that the biological signatures of the hybrid 2D-QSAR models prevailed in fact in all of them. Out of eight descriptors of these hybrid models, 6–7 descriptors belong to the biological signatures, clearly pinpointing their key role. Furthermore, some biological signatures like D0253 and D2942 appear almost in every model, whereas D0406, D1581 and D2035 were found to be present in multiple models.

As shown in Fig. 5.3, the relative significance of the descriptors computed for each of these four models patently portrays the fact that it is significantly lower for the chemical descriptors than for the biological signatures. Therefore, biological signatures mainly prevailed in these hybrid models, and even if with low importance helped in improving the quality of the models to a considerable extent.

The determined Williams plots are shown in Fig. 5.4. Inspection of these plots reveals that, save for Model 3 developed by considering both SIRMS and signaturizers descriptors, no structural outliers were detected in these models. It may therefore be inferred that the descriptors of these models assured that the data-points are within their AD. Moreover, two structural outliers of Model 2 were in fact predicted very well by this model. All hybrid models contain two response outliers which naturally

Table 5.6 Detailed outline of the four most predictive linear 2D-QSAR models obtained in the present work by mixing different type of descriptors

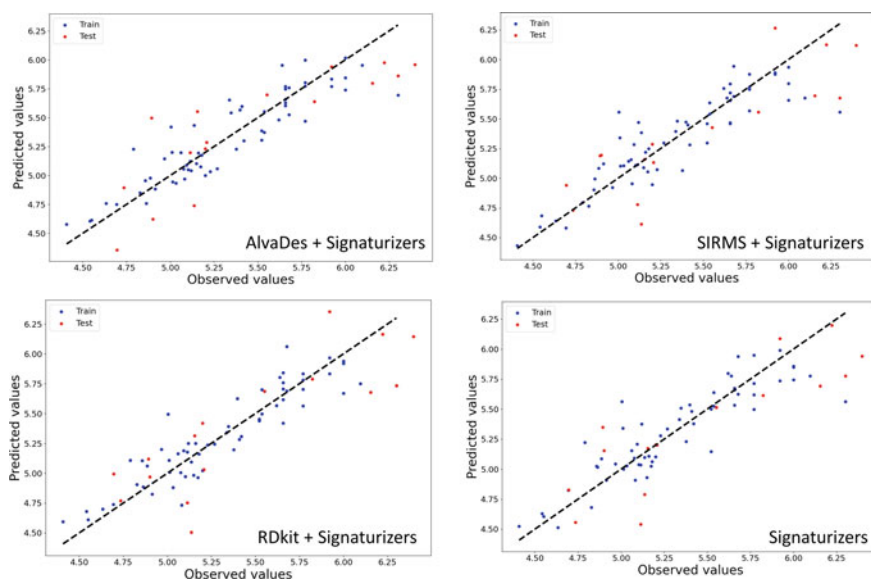
Model	Descriptors	Equation	Statistical results ^a
1	AlvaDesc + signaturizers	$pEC_{50} = + 3.728 (\pm 0.205) - 3.000 (\pm 0.602) D0253 + 1.597 (\pm 0.609) D0791 - 6.828 (\pm 0.940) D1649 + 1.858 (\pm 0.544) D2144 - 3.717 (\pm 0.898) D2830 - 4.380 (\pm 0.648) D2942 + 0.092 (\pm 0.018) CATS3D_{09_DL} + 0.143 (\pm 0.029) CATS3D_{08_PL}$	$N_{\text{training}} = 61; R^2 = 0.821; R^2_{\text{Adj}} = 0.790; F(52; 8) = 29.780; Q^2_{\text{LOO}} = 0.743; MAE = 0.137; MSE = 0.179; r_m^2_{\text{LOO}} = 0.668; \Delta r_m^2_{\text{LOO}} = 0.153; N_{\text{test}} = 16; R^2_{\text{Pred}} = 0.720; RMSEP = 0.311; r_m^2_{\text{test}} = 0.631; \Delta r_m^2_{\text{test}} = 0.043; R^2_{\text{Test}} = 0.719; k = 1.013, k' = 0.984; r_0^2 - r'_0{}^2 = 0.078; cRp^2 = 0.756$
2	SIRMS + signaturizers	$pEC_{50} = + 5.767 (\pm 0.151) - 2.918 (\pm 0.586) D0253 - 1.412 (\pm 0.521) D0856 + 4.232 (\pm 0.677) D1581 + 5.338 (\pm 1.480) D2035 + 1.525 (\pm 0.521) D2199 + 0.816 (\pm 0.589) D2492 - 3.768 (\pm 0.673) D2942 - 0.321 (\pm 0.073) S n 4 REFRACTIVITY C-D.C = D$	$N_{\text{training}} = 61; R^2 = 0.791; R^2_{\text{Adj}} = 0.759; F(52; 8) = 24.569; Q^2_{\text{LOO}} = 0.741; MAE = 0.136; MSE = 0.194; r_m^2_{\text{LOO}} = 0.644; \Delta r_m^2_{\text{LOO}} = 0.068; N_{\text{test}} = 16; R^2_{\text{Pred}} = 0.722; RMSEP = 0.309; r_m^2_{\text{test}} = 0.648; \Delta r_m^2_{\text{test}} = 0.020; R^2_{\text{Test}} = 0.733; k = 1.017, k' = 0.980; r_0^2 - r'_0{}^2 = 0.079; cRp^2 = 0.724$
3	RDKit + signaturizers	$pEC_{50} = + 4.967 (\pm 0.219) - 3.311 (\pm 0.449) D0253 - 3.260 (\pm 0.580) D0406 + 3.071 (\pm 0.733) D0448 + 4.523 (\pm 0.653) D1581 + 5.914 (\pm 1.360) D2035 - 4.830 (\pm 0.628) D2942 + 0.314 (\pm 0.116) MoRSE42 + 0.006 (\pm 0.001) RDF197$	$N_{\text{training}} = 61; R^2 = 0.821; R^2_{\text{Adj}} = 0.794; F(52; 8) = 29.829; Q^2_{\text{LOO}} = 0.742; MAE = 0.135; MSE = 0.179; r_m^2_{\text{LOO}} = 0.647; \Delta r_m^2_{\text{LOO}} = 0.142; N_{\text{test}} = 16; R^2_{\text{Pred}} = 0.710; RMSEP = 0.316; r_m^2_{\text{test}} = 0.626; \Delta r_m^2_{\text{test}} = 0.032; R^2_{\text{Test}} = 0.714; k = 1.010; k' = 0.987; r_0^2 - r'_0{}^2 = 0.027; cRp^2 = 0.756$

(continued)

Table 5.6 (continued)

Model	Descriptors	Equation	Statistical results ^a
4	Signaturizers	$pEC_{50} = 5.578 (\pm 0.176) - 3.1407 (\pm 0.522) D0253 - 2.440 (\pm 0.597) D0406 + 1.526 (\pm 0.533) D0605 + 3.763 (\pm 0.675) D1581 + 6.832 (\pm 1.4912) D2035 - 1.795 (\pm 0.631) D2619 + 1.601 (\pm 0.675) D2895 - 4.444 (\pm 0.689) D2942$	$N_{\text{training}} = 61; R^2 = 0.788; R^2_{\text{Adj}} = 0.756; F(52; 8) = 24.210; Q^2_{\text{LOO}} = 0.717; MAE = 0.142; MSE = 0.195; r_m^2_{\text{LOO}} = 0.615; \Delta r_m^2_{\text{LOO}} = 0.147; N_{\text{test}} = 16; R^2_{\text{Pred}} = 0.720; RMSEP = 0.311; r_m^2_{\text{test}} = 0.660; \Delta r_m^2_{\text{test}} = 0.030; R^2_{\text{Test}} = 0.742; k = 1.020; k' = 0.977; r_0^2 - r'^2_0 = 0.064; cRp^2 = 0.723$

^a N_{training} : number of training set compounds; R^2 : determination coefficient; R^2_{Adj} : adjusted R^2 ; F : Fisher statistics; Q^2_{LOO} : leave-one-out cross-validated R^2 ; MAE: mean absolute error; MSE: mean square error; $r_m^2_{\text{LOO}}$: leave-one-out r_m^2 metric; $\Delta r_m^2_{\text{LOO}}$: standard deviation of $r_m^2_{\text{LOO}}$; N_{test} : number of test set compounds; R^2_{Pred} : R^2 for external prediction; RMSEP: root mean square error of prediction; $r_m^2_{\text{test}}$: r_m^2 for the test set; $\Delta r_m^2_{\text{test}}$: standard deviation of $r_m^2_{\text{test}}$; R^2_{Test} , k , k' and $|r_0^2 - r'^2_0|$: parameters belong to Golbraikh and Tropsha's acceptable model criteria for test set validation; and cRp^2 : statistical parameter of the Y -randomization test [28–31, 33]

**Fig. 5.2** Observed versus predicted plots of the 2D-QSAR hybrid models

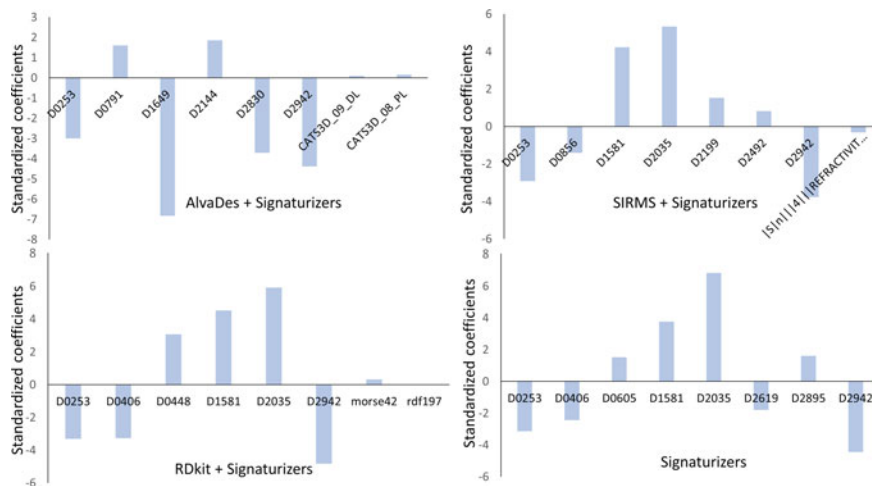


Fig. 5.3 Relative significance of the descriptors in the 2D-QSAR hybrid models

lowers their overall predictivity, but these should not be removed since they lie very well within the applicability structural domain of the models.

We finally left with the question whether non-linear 2D-QSAR models with higher statistical predictivity might exist using the employed descriptors so far. Even though

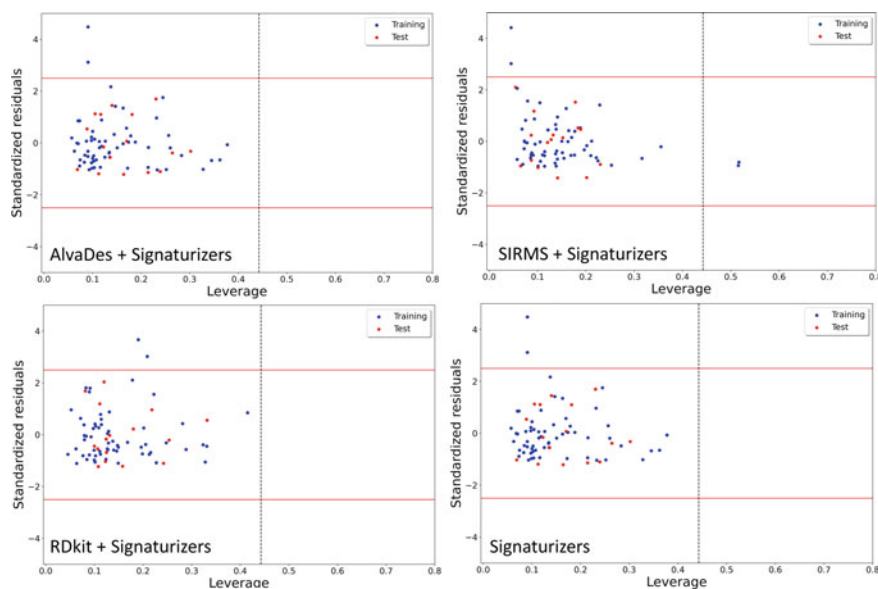


Fig. 5.4 Williams plots of the 2D-QSAR models

the main goal of this work is to develop linear models, since most of the chemical descriptors failed to accomplish such a goal, the question becomes even more significant. Therefore, we also attempted to develop non-linear models to check that. Nowadays, it is well known that there are multiple ways of developing non-linear models as regards to (a) the applied machine learning (ML) technique, (b) the setting of parameters for the targeted ML technique and (c) the descriptors selection strategy. In this work, we paid particular attention on two different schemes for developing such non-linear models using the following three ML techniques: (a) support vector regression (SVR), (b) random forests regression (RFR) and (c) multilayer perception-based regression (MLPR). The models were firstly developed after careful hyperparameter optimisation, the details of which were provided in Table 5.2. As far as the descriptor selection is concerned, we followed two schemes. In the first one, descriptors and data distributions obtained from the best linear models were considered for model development. In the second, 20 most distinct descriptors were obtained from the differential Shannon entropy (dSe) technique calculated with the IMMAN software (<http://mobiosd-hub.com/imman-soft/>) [36, 37]. A statistical summary of the performance of the best non-linear models found by following these schemes is given in Table 5.7.

The attained results may be summarised as follows:

- (a) Descriptors selected by dSe as well the descriptors selected directly from the linear 2D-QSAR models failed to generate predictive non-linear models. However, the performance of the models was better when the descriptors of the respective linear model were deployed for model generation. Therefore, we did not consider the dSe selected descriptors for deriving hybrid non-linear models, that is, based on chemical descriptors plus bioactivity descriptors.
- (b) Even more importantly, none of the non-linear models achieved a statistical predictivity significantly higher than that pertaining to the linear 2D-QSAR models.
- (c) The SVR remained the most successful regressor among the three ML techniques employed for model generation. All predictive SVR models were derived with a 'linear' kernel and not with a 'RBF' kernel. Nevertheless, the SVR models were not statistically more predictive when compared to the linear 2D-QSAR models (see Table 5.5).
- (d) Hybrid non-linear models were found to be more predictive compared to models generated either with only chemical descriptors or with only biological signatures. The best non-linear model found was produced by SVM with AlvaDesc and signaturizer descriptors ($Q^2_{\text{LOO}} = 0.750$, $R^2_{\text{Pred}} = 0.705$). However, the statistical quality of this model was no better than that of the linear models.
- (e) Finally, it is worth mentioning here that we even attempted to derive models including all the descriptors from each set, after removing the constant and near-constant descriptors as well as highly correlated features by setting the correlation cut-off to 0.95 and the variance cut-off to 0.001. Still, we realised that the quality of the non-linear models rather deteriorates with the increase in the number of descriptors (results not shown).

Table 5.7 Summary of the statistical results achieved for the non-linear models

Descriptors	Random seed	Type of descriptors	ML	Q^2_{LOO}	R^2_{Pred}	Average ^a
AlvaDesc	42	dSe	MLPR	- 0.463	0.522	0.030
AlvaDesc	42	dSe	RFR	0.203	0.258	0.231
AlvaDesc	42	dSe	SVR	0.198	0.182	0.190
RDKit	3	dSe	MLPR	- 0.049	0.028	- 0.011
RDKit	3	dSe	RFR	0.264	0.135	0.200
RDKit	3	dSe	SVR	0.140	0.162	0.151
SIRMS	3	dSe	MLPR	0.169	- 0.073	0.048
SIRMS	3	dSe	RFR	0.384	0.209	0.297
SIRMS	3	dSe	SVR	0.384	0.209	0.297
Signaturizers	20	dSe	MLPR	0.183	0.176	0.180
Signaturizers	20	dSe	RFR	0.110	0.237	0.174
Signaturizers	20	dSe	SVR	0.162	0.315	0.239
AlvaDesc	42	Linear model	MLPR	0.190	0.526	0.358
AlvaDesc	42	Linear model	RFR	0.426	0.432	0.429
AlvaDesc	42	Linear model	SVR	0.579	0.556	0.568
RDKit	3	Linear model	MLPR	0.003	0.099	0.051
RDKit	3	Linear model	RFR	0.317	0.522	0.420
RDKit	3	Linear model	SVR	0.655	0.516	0.586
SIRMS	3	Linear model	MLPR	0.427	- 0.066	0.181
SIRMS	3	Linear model	RFR	0.472	0.514	0.493
SIRMS	3	Linear model	SVR	0.310	0.513	0.412
Signaturizers	20	Linear model	MLPR	0.200	0.163	0.182
Signaturizers	20	Linear model	RFR	0.391	0.536	0.464
Signaturizers	20	Linear model	SVR	0.715	0.695	0.705
AlvaDesc + Signaturizers	20	Linear model	MLPR	0.362	0.598	0.480

(continued)

Table 5.7 (continued)

Descriptors	Random seed	Type of descriptors	ML	Q^2_{LOO}	R^2_{Pred}	Average ^a
AlvaDesc + Signaturizers	20	Linear model	RFR	0.470	0.450	0.460
AlvaDesc + Signaturizers	20	Linear model	SVR	0.750	0.705	0.728
RDKit + Signaturizers	20	Linear model	MLPR	- 0.200	0.132	- 0.034
RDKit + Signaturizers	20	Linear model	RFR	0.592	0.713	0.653
RDKit + Signaturizers	20	Linear model	SVR	0.695	0.669	0.682
SIRMS + Signaturizers	20	Linear model	MLPR	0.203	0.210	0.207
SIRMS + Signaturizers	20	Linear model	RFR	0.603	0.180	0.392
SIRMS + Signaturizers	20	Linear model	SVR	- 0.147	- 0.121	- 0.134

^aAverage value of Q^2_{LOO} and R^2_{Pred}

5.4 Conclusions

In this chapter, we attempted to highlight the importance of the newly developed descriptors—i.e., bioactivity descriptors or biological signatures, for setting up predictive 2D-QSAR models. Such descriptors require only the SMILES notation for the targeted compounds and provide a range of descriptor values jointly embodying their chemical, biological and clinical profiles. As a case study for such purpose, we employed a dataset comprising 77 compounds with cell-based biological activity against the DENV-2 protease. What is more, it is also important to understand the significance of the current work from the context of the nature of biological activity data used for modelling. As referred to earlier, the outcomes of such cell-based assays (i.e., the DENV2ProHeLa assay) are influenced not only by the type of biological target (i.e., the NS2B-NS3 protease) but also by the complex multifactorial conditions that do exist inside a specific cellular system. The less satisfactory performance of chemical descriptors to characterise the structure activity relationships may well be explained from the fact that they fail to encode the complexity of biological results. Therefore, most likely the outcomes of such cell-based assays can only be modelled by some kind of bioactivity descriptors that not only encode chemical attributes but also biological profiles with numerical values. From this very reason, we were encouraged to explore the newly developed bioactivity descriptors (also named signaturizers) for building 2D-QSAR models. For the sake of comparisons, we attempted to develop linear 2D-QSAR models using different sets of chemical descriptors calculated with a number of different software packages/programs. As

such, the present work paid particular attention to both the robustness and consistency of the models' development techniques. For each descriptor, we generated as many as 44 (i.e., 24 SFS-MLR plus 20 GA-MLR) different models to select the most predictive model, which is reported here. Regardless of chemical descriptors (or fragments) or model development techniques, no linear 2D-QSAR model was found with satisfactory statistical predictivity. The global signaturizer descriptors however supplied us a linear 2D-QSAR model, the overall statistical quality of which was around 20% better than the most predictive model generated with chemical descriptors. Bioactivity descriptors were then merged with chemical descriptors to generate hybrid linear models in a bid to improve the overall predictivities of the 2D-QSAR models. The combination of signaturizers with AlvaDesc descriptors afforded the most predictive linear hybrid model, although SIRMS and RDKit descriptors also delivered hybrid models with similar statistical predictivity. What is more, non-linear models generated with multiple machine learning techniques also showed the importance of bioactivity descriptors. The results from this work therefore mean that the newly proposed biological signatures proposed by Bertoni et al. [13] shall be very useful in the future for developing predictive 2D-QSAR models. Naturally, their true significance may only be established when these bioactivity descriptors are compared with other chemical descriptors just as it was carried out here. Thanks to that, this work conveys important guidelines to exploit different linear and non-linear model development strategies in a systematic and consistent manner. The entire work outlined in this chapter is based on non-commercial open-access tools, programs and webservers, so that the models can easily be reproduced, and a model development strategic landscape followed in the future as well.

Acknowledgements This work was supported by UIDB/50006/2020 with funding from FCT/MCTES through national funds.

References

1. Hasan S, Jamdar S, Alalawi M, Al Ageel Al Beajji S (2016) *J Int Soc Prev Community Dent* 6:1–6. <https://doi.org/10.4103/2231-0762.175416>
2. Wang W-H, Urbina AN, Chang MR, Assavalapsakul W, Lu P-L, Chen Y-H, Wang S-F (2020) *J Microbiol Immunol Infect* 53:963–978. <https://doi.org/10.1016/j.jmii.2020.03.007>
3. Aranda, C, Martínez MJ, Montalvo T, Eritja R, Navero-Castillejos J, Herreros E, Marqués E, Escosa R, Corbella I, Bigas E, Picart L, Jané M, Barrabeig I, Torner N, Talavera S, Vázquez A, Sánchez-Seco MP, Busquets N (2018) *Euro Surveill* 23:1700837. <https://doi.org/10.2807/1560-7917.Es.2018.23.47.1700837>
4. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, Drake JM, Brownstein JS, Hoen AG, Sankoh O, Myers MF, George DB, Jaenisch T, Wint GRW, Simmons CP, Scott TW, Farrar JJ, Hay SI (2013) *Nature* 496:504–507. <https://doi.org/10.1038/nature12060>
5. Ojha A, Ni D, Batra H, Singhal R, Annarapu GK, Bhattacharyya S, Seth T, Dar L, Medigeshi GR, Vrati S, Vikram NK, Guchhait P (2017) *Sci Rep* 7:41697. <https://doi.org/10.1038/srep41697>
6. Kuo H-J, Lee I-K, Liu J-W (2018) *J Microbiol Immunol Infect* 51:740–748. <https://doi.org/10.1016/j.jmii.2016.08.024>

7. Guzman MG, Halstead SB, Artsob H, Buchy P, Farrar J, Gubler DJ, Hunsperger E, Kroeger A, Margolis HS, Martínez E, Nathan MB, Pelegrino JL, Simmons C, Yoksan S, Peeling RW (2010) *Nat Rev Microbiol* 8:S7–S16. <https://doi.org/10.1038/nrmicro2460>
8. Kuhl N, Leuthold MM, Behnam MAM, Klein CD (2021) *J Med Chem* 64:4567–4587. <https://doi.org/10.1021/acs.jmedchem.0c02042>
9. Normile D (2017) *Science* 358:1514–1515. <https://doi.org/10.1126/science.358.6370.1514>
10. Thomas SJ, Yoon I-K (2019) *Hum Vaccines Immunother* 15:2295–2314. <https://doi.org/10.1080/21645515.2019.1658503>
11. Barrows NJ, Campos RK, Liao K-C, Prasanth KR, Soto-Acosta R, Yeh S-C, Schott-Lerner G, Pompon J, Sessions OM, Bradrick SS, Garcia-Blanco MA (2018) *Chem Rev* 118:4448–4482. <https://doi.org/10.1021/acs.chemrev.7b00719>
12. Kühl N, Graf D, Bock J, Behnam MAM, Leuthold M-M, Klein CD (2020) *J Med Chem* 63:8179–8197. <https://doi.org/10.1021/acs.jmedchem.0c00413>
13. Bertoni M, Duran-Frigola M, Badia-i-Mompel P, Pauls E, Orozco-Ruiz M, Guitart-Pla O, Alcalde V, Diaz VM, Berenguer-Llgero A, Brun-Heath I, Villegas N, de Herrerros AG (2021) *Nat Commun* 12:3932. <https://doi.org/10.1038/s41467-021-24150-4>
14. Duran-Frigola M, Pauls E, Guitart-Pla O, Bertoni M, Alcalde V, Amat D, Juan-Blanco T, Aloy P (2020) *Nat Biotechnol* 38:1087–1096. <https://doi.org/10.1038/s41587-020-0502-7>
15. Sushko I, Novotarskyi S, Korner R, Pey AK, Rupp M, Teetz W, Brmaier S, Abdelaziz A, Prokopenko VV, Tanchuk VY, Todeschini R, Varnek A, Marcou G, Ertl P, Potemkin V, Grishina M, Gasteiger J, Schwab C, Baskin II, Palyulin VA, Radchenko EV, Welsh WJ, Kholodovych V, Chekmarev D, Cherkasov A, Aires-de-Sousa J, Zhang QY, Bender A, Nigsch F, Patiny L, Williams A, Tkachenko V, Tetko IV (2011) *J Comput Aided Mol Des* 25:533–554. <https://doi.org/10.1007/s10822-011-9440-2>
16. Mauri A (2020) Ecotoxicological QSARs. In: Roy K (ed) *Methods in pharmacology and toxicology*. Humana, New York, NY, pp 801–820. https://doi.org/10.1007/978-1-0716-0150-1_32
17. de Sousa JMA (2017) In: Varnek A (ed) *Tutorials in chemoinformatics*, pp 127–134. <https://doi.org/10.1002/9781119161110.ch8>
18. Kuz'min VE, Artemenko AG, Polischuk PG, Muratov EN, Hromov AI, Liahovskiy AV, Andronati SA, Makan SY (2005) *J Mol Model* 11:457–467. <https://doi.org/10.1007/s00894-005-0237-x>
19. Varnek A, Fourches D, Hoonakker F, Solov'ev VP (2005) *J Comput-Aided Mol Des* 19:693–703. <https://doi.org/10.1007/s10822-005-9008-0>
20. Filimonov D, Poroikov V, Borodina Y, Glorizova T (1999) *J Chem Inf Comput Sci* 39:666–670. <https://doi.org/10.1021/ci980335o>
21. Potemkin VA, Grishina MA, Bartashevich EV (2007) *J Struct Chem* 48:155–160. <https://doi.org/10.1007/s10947-007-0023-y>
22. Moriwaki H, Tian Y-S, Kawashita N, Takagi T (2018) *J Cheminform* 10:4. <https://doi.org/10.1186/s13321-018-0258-y>
23. Masand VH, Rastija V (2017) *Chemom Intell Lab Syst* 169:12–18. <https://doi.org/10.1016/j.chemolab.2017.08.003>
24. ChemAxon (2010) *Standardizer*, Budapest HCVS
25. Halder AK, Delgado AHS, Cordeiro MNDS (2022) *Dent Mater* 38:333–346. <https://doi.org/10.1016/j.dental.2021.12.014>
26. Ambure P, Aher RB, Gajewicz A, Puzyr T, Roy K (2015) *Chemom Intell Lab Syst* 147:1–13. <https://doi.org/10.1016/j.chemolab.2015.07.007>
27. Halder AK, Cordeiro MNDS (2021) *J Cheminform* 13:29. <https://doi.org/10.1186/s13321-021-00508-0>
28. Tetko IV, Tanchuk VY, Villa AEP (2001) *J Chem Inf Comput Sci* 41:1407–1421. <https://doi.org/10.1021/ci010368v>
29. Roy PP, Paul S, Mitra I, Roy K (2009) *Molecules* 14:1660–1701. <https://doi.org/10.3390/molcules14051660>

30. Golbraikh A, Tropsha A (2002) *J Mol Graph Model* 20:269–276. [https://doi.org/10.1016/S1093-3263\(01\)00123-1](https://doi.org/10.1016/S1093-3263(01)00123-1)
31. Vrontaki E, Melagraki G, Mavromoustakos T, Afantitis A (2015) *Methods* 71:4–13. <https://doi.org/10.1016/j.ymeth.2014.03.021>
32. Yoo W, Mayberry R, Bae S, Singh K, Peter He Q, Lillard JW Jr (2014) *Int J Appl Sci Technol* 4:9–19
33. Ojha PK, Roy K (2011) *Chemom Intell Lab Syst* 109:146–161. <https://doi.org/10.1016/j.chemolab.2011.08.007>
34. Serra A, Önlü S, Festa P, Fortino V, Greco D, Ponty Y (2020) *Bioinformatics* 36:145–153. <https://doi.org/10.1093/bioinformatics/btz521>
35. Gramatica P (2007) *QSAR Comb Sci* 26:694–701. <https://doi.org/10.1002/qsar.200610151>
36. Urias RWP, Barigye SJ, Marrero-Ponce Y, García-Jacas CR, Valdes-Martíni JR, Perez-Gimenez F (2015) *Mol Divers* 19:305–319. <https://doi.org/10.1007/s11030-014-9565-z>
37. Speck-Planche A (2019) *ACS Omega* 4:3122–3132. <https://doi.org/10.1021/acsomega.8b03693>

Part III
SMILES for QSPR/QSAR with Optimal
Descriptors

Chapter 6

QSPR Models for Prediction of Redox Potentials Using Optimal Descriptors



Karel Nesměrák and Andrey A. Toropov

Abstract The redox potential is an important physicochemical property widely used for the characterization of chemical species, and, as a characteristic constant of a given chemical species, it is also useful for predicting various other properties of the species. In the chapter, we review and discuss the pros and cons of QSPR models for the prediction of redox potentials using optimal descriptors calculated with the SMILES as well as using the so-called hybrid descriptors calculated with considering SMILES and molecular graphs of atomic orbitals.

Keywords QSPR · Redox potential · Drug design · Monte Carlo method

6.1 Introduction, Redox Potential, and Its Significance

The electron and its transfer play a fundamental role in chemical reactions, processes that are very common in our real world and that lead to the chemical transformation of one set of chemical substances to another [1]. When a chemical reaction involves a change in the oxidation states of the reactants, we refer to such a reaction as an oxidation–reduction reaction, or redox reaction, for short. A reactant that has a strong affinity for electrons (an electron acceptor) is referred to as an oxidant, and its oxidation number decreases during the reaction. The opposite is a reactant called a reductant, which is an electron donor, and its oxidation number increases during the reaction.

The tendency of a chemical species to electron transfer is characterized by the oxidation–reduction (redox) potential [2]. The redox reaction between two

K. Nesměrák (✉)

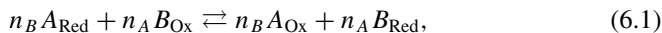
Department of Analytical Chemistry, Faculty of Science, Charles University, Hlavova 8, 128 43 Prague 2, Czech Republic

e-mail: karel.nesmerak@natur.cuni.cz

A. A. Toropov

Laboratory of Environmental Chemistry and Toxicology, Department of Environmental Health Science, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156 Milano, Italy

substances can be represented by the general chemical equation,



where $A_{\text{Ox/Red}}$, respectively, $B_{\text{Ox/Red}}$, is termed oxidation–reduction chemical pair, and n_A , respectively, n_B , is the stoichiometric coefficient. Equation (6.1) also represents the chemical transformation of the system from state I to state II. The equilibrium constant based on the activities a_i of the individual reactants or products can be written for this reaction as

$$K = \frac{a_{A_{\text{Ox}}}^{n_B} a_{B_{\text{Red}}}^{n_A}}{a_{A_{\text{Red}}}^{n_B} a_{B_{\text{Ox}}}^{n_A}}. \quad (6.2)$$

The shift of equilibrium of any chemical reaction depends, at constant temperature and pressure, on the change in the free enthalpy ΔG corresponding to the transition from state I to state II, which can be expressed as the change of the chemical potential μ_i , that is for Eq. (6.1)

$$\Delta G = G_{\text{II}} - G_{\text{I}} = n_B \mu_{A_{\text{Ox}}} + n_A \mu_{B_{\text{Red}}} - n_B \mu_{A_{\text{Red}}} - n_A \mu_{B_{\text{Ox}}}. \quad (6.3)$$

The chemical potential is generally defined on the basis of the activity of a substance by the relationship,

$$\mu_i = \mu_i^\circ + RT \ln a_i, \quad (6.4)$$

where μ_i° is standard chemical potential, R is the molar gas constant (8.314 J K⁻¹ mol⁻¹), and T is thermodynamic temperature.

Inserting Eq. (6.4) into Eq. (6.3) and rearranging leads to

$$\begin{aligned} \Delta G &= n_B \mu_{A_{\text{Ox}}}^\circ + n_A \mu_{B_{\text{Red}}}^\circ - n_B \mu_{A_{\text{Red}}}^\circ - n_A \mu_{B_{\text{Ox}}}^\circ + RT \ln \frac{a_{A_{\text{Ox}}}^{n_B} a_{B_{\text{Red}}}^{n_A}}{a_{A_{\text{Red}}}^{n_B} a_{B_{\text{Ox}}}^{n_A}} \\ &= \Delta G^\circ + RT \ln \frac{a_{A_{\text{Ox}}}^{n_B} a_{B_{\text{Red}}}^{n_A}}{a_{A_{\text{Red}}}^{n_B} a_{B_{\text{Ox}}}^{n_A}}. \end{aligned} \quad (6.5)$$

At the same time, the redox reaction can be seen as a chemical work in which a certain number of electrons n is transferred and the potential difference E between the two redox pairs is overcome [3, 4]. The change in free enthalpy representing this chemical work is given,

$$\Delta G = -nFE, \quad (6.6)$$

where F is the Faraday constant (96,485 C mol⁻¹), representing the electric charge of one mole of electrons.

Substituting Eq. (6.6) into Eq. (6.5), we obtain the famous Nernst equation, which represents the basic relation defining the dependence of the equilibrium redox potential on the activity of the electroactive species,

$$E = E^\circ - \frac{RT}{nF} \ln \frac{a_{A_{Ox}}^{n_B} a_{B_{Red}}^{n_A}}{a_{A_{Red}}^{n_B} a_{B_{Ox}}^{n_A}}, \quad (6.7)$$

in which E° is a characteristic of the redox system (called the standard redox potential) that is intrinsically linked to the chemical nature of the species, and the term after the logarithm describes the effect of the actual composition of the system.

When the system reaches equilibrium ($\Delta G = 0$), Eq. (6.7) goes – applying simultaneously Eq. (6.2) – to the form,

$$E^\circ = \frac{RT}{nF} \ln K, \quad (6.8)$$

which is the fundamental relationship between the standard redox potential and the equilibrium constant. By combining Eq. (6.8) with Eq. (6.6), we obtain the relationship between the standard redox potential and the standard Gibbs energy of the system,

$$\Delta G^\circ = -RT \ln K = -nFE^\circ. \quad (6.9)$$

This equation describes the intrinsic relationship between the change in free energy for a chemical reaction and the redox potential value. In addition, as will be shown below, this is the basic relationship underlying the possibility of a correlation between the structure and redox potential.

The standard redox potential is a characteristic constant for a given molecule/species and is directly connected to its chemical structure. It is an important physicochemical characteristic of any chemical species that characterizes the ease, or difficulty, of structural changes of this molecule related to the transfer of electrons. Therefore, the redox potential is of great importance both for chemistry as such and for the application of chemical species in biological systems in which redox reactions are predominant (hence its application in medicinal chemistry, e.g., in drug development) [5, 6]. The redox potential also finds application in many other areas of applied chemistry and chemical technology [7, 8].

Like other physicochemical quantities, the redox potential can be obtained experimentally, based on electrochemical measurements [2]. In particular, wide ranges of voltammetric techniques are used in which a signal, which results from the interaction of electrons directly with the chemical species under study, is obtained. In addition, the facile variability of the electrochemical measurement conditions makes it easy to change the desired conditions for the reaction under investigation, for example, by measuring at different pH [9] or in the non-aqueous medium [10, 11]. Depending on the measurement technique used, the redox potential may be expressed as the half-wave potential ($E_{1/2}$) measured by direct-current voltammetry or the peak

potential (E_p) measured by, most often, differential pulse voltammetry. Using cyclic voltammetry, the potentials of the oxidation (E_{pa}) or reduction (E_{pc}) peaks are available. In some cases, calculated values such as $E_{redox} = (E_{pc} + E_{pa})/2$ (for reversible systems) or $E_{mid} = E_p - E_{p/2}$ (for irreversible systems) are used. Occasionally, other parameters are used [12, 13].

6.2 Relationship Between Redox Potential and Structure

The wide application of the redox potential is the reason why quantitative relationships between the structure of a chemical species and the value of the redox potential are sought [14]. The possibility of quantifying the relationships between structure and redox potential leads to a better understanding of the role of redox properties of a chemical species in its chemical, biological, therapeutic, or other action. This can be found using methods of quantitative structure–activity/structure–property relationships (QSAR/QSPR). The objective of QSPR is to find a function, described by a mathematical equation, of the dependence of physicochemical property on the structure of a chemical species [15–17].

QSPR allows general conclusions to be drawn from experimental data and to predict the behavior and properties of unstudied or even non-existent chemical species. This is a practical application of the central assumption of QSPR that structurally similar molecules have similar properties [18]. A tool to achieve the objectives of QSPR is to compare quantitative experimental and theoretical data using different mathematical models and procedures. In short, the result of any QSPR should be the equation

$$\text{Endpoint} = \text{mathematical function (Molecular descriptors)}, \quad (6.10)$$

where molecular descriptors are a set of calculated or measured values that effectively describe the molecular structure of a chemical species.

The process of producing QSPR models essentially follows the procedures used in any conventional data mining task. Thus, the process consists of five basic steps [19]:

1. Measurement of physicochemical data and their processing.
2. Selection and calculation or measurement of appropriate molecular descriptors.
3. Model establishing and training.
4. Model validation.
5. Determination of the applicability of the QSPR model.

Generally, physicochemical data are the most important component of any QSPR. Many studies indicated that the quantity and quality of input physicochemical data seriously affect the quality of the model [20, 21]. When dealing with experimental data, the study is always dependent on the data provider, and care should be taken to be aware of errors and variability/irregularities in the data [22].

Molecular descriptors play a crucial role in any QSPR. They are derived using graph theory, information theory or physical, quantum, and organic chemistry. Thousands of types of molecular descriptors are currently defined [23]. Molecular descriptors can be roughly divided into two main groups:

1. Descriptors based on experimental measurements (e.g., octanol–water partition coefficient, molar refractivity), which are generally applicable as physicochemical descriptors.
2. Theoretical descriptors, which are derived from symbolic representations of the molecule (e.g., graph theory) or are derived from physicochemical theories and have some natural overlap with experimental methods (e.g., Hammett constants).

The fundamental difference between experimental and theoretical molecular descriptors is that theoretical descriptors, unlike experimental ones, do not contain statistical error due to noise in experimental measurements.

The relationship between the redox potential of chemical species and its structure was already noticed by one of the founders of modern electrochemistry, Heyrovsky, who in 1934 defined the conjugation rule [24]: ‘The polarographic reduction becomes easier as the number of conjugated bonds in the organic molecule increases.’ The next empirical rule was the electronegativity rule formulated in 1938 by Shikata and Tachi [25]: ‘The more electronegative the substituent, the more positive the half-wave potential.’

The actual quantification of the relationship between redox potential and chemical species structure was only possible after the introduction of the Hammett approach to QSPR. Hammett studied the effect of substituents on the reaction rate constants of a series of substituted organic acids [26]. From the results, he postulated that the effect of substitution (i.e., the change in the distribution of electrons in a compound due to a substituent) on the quantitative change in a property (the value of the rate or equilibrium constant) could be expressed by the equation,

$$\log k_X = \log k_H + \rho\sigma_X, \quad (6.11)$$

where k_X is the rate (or equilibrium) constant of the substituted derivative, k_H is the rate (equilibrium) constant of the unsubstituted derivative (with a hydrogen atom in place of the substituent), ρ is the reaction constant, which is a measure of the sensitivity of a given reaction to the electronic effect of substituents (and is therefore characteristic of the reaction), and σ_X is the Hammett constant of the substituent, describing—in general, since it is transferable between single reactions—the effect of the substituent on the distribution of electrons in a given molecule. Equation (6.11) became one of the first examples of the approach that received the name linear free energy (Gibbs energy) relationship [18], and Hammett constant became the first descriptor that allowed the encoding of chemical information into a mathematical expression.

By combining Eq. (6.11) with Eq. (6.8), the Hammett equation for the redox potential is obtained,

$$E_X^\circ = E_H^\circ + \rho\sigma_X, \quad (6.12)$$

where E_X° is the standard redox potential of the substituted derivative and E_H° is the standard redox potential of the unsubstituted derivative.

Currently, a variety of techniques is used for QSPR relationships describing the effect of structure on the redox potential of a chemical speciation. The fundamental work on QSPR of the redox potential was published by Zuman [27]. In the 1990s, the subfield of QSPR that deals with the influence of the structure of a chemical species on its electrochemical properties acquired the acronym QSER, that is, quantitative structure–electrochemical relationships [28]. Table 6.1 summarizes recent QSER relationships between a redox potential and a structure.

Table 6.1 A review of recently published papers on quantitative structure–electrochemical relationships for redox potential using different structural descriptors (descriptors used class of compounds, number of compounds in study, squared correlation coefficient of test set, references)

Descriptors	Compounds	Number of compounds	R^2	References
Electronic effect descriptor	1,4-Naphthoquinones	19	0.96	[29]
Electrophilicity index	Quinones	26	0.98	[30]
Group of different descriptors	Chlorinated organic compounds	21	0.88	[31, 32]
	Quinones	36	< 0.36	[33]
	Steroids	40	n/a	[34]
Hammett constants	9-Anilinoacridines	18	0.69	[35]
	1,4-Benzoquinones	54	0.79	[28]
	Benzoxazines	40	0.90	[36]
	Benzylideneanilines	49	0.89	[37]
	4-(Benzylsulfanyl)pyridines	22	0.99	[38]
	1,4-Naphthoquinones	30	0.83	[28]
	Polysubstituted benzenes	9	n/a	[39]
	α,β -Unsaturated ketones	17	0.98	[40]
	α,β -Unsaturated ketones	11	0.99	[41]
Minimum charges on oxygen atoms	Quinones	9	n/a	[42]
Molecular graphs	Aldehydes and ketones	73	> 0.80	[43]
	Anthraquinones	30	0.96	[44]
	Anthraquinones	33	0.94	[45]

(continued)

Table 6.1 (continued)

Descriptors	Compounds	Number of compounds	R^2	References
Molecular orbital energy	Steroids	38	0.59	[46]
	Benzoxazines	40	0.80	[36]
	Flavonoids	29	0.93	[47]
	Polycyclic aromatic hydrocarbons	44	0.99	[48]
Polarizability ZZ index	Carotenoids	23	0.77	[49]
Quantum chemical	Benzoxazines	40	0.95	[50]
	Benzylsulfanyltetrazaoles	19	0.98	[51]
	Nitrobenzenes	15	0.96	[52]
	Phenylquinolinylethynes	30	0.84	[53]
	Quinones	8	n/a	[54]
	Quinones	10	n/a	[55, 56]
	Quinones	18	n/a	[57]
	Quinones	5	n/a	[58]
	Squaric acid	5	n/a	[59]
	Thioxanthenes	4	n/a	[60]
Swain–Lupton	1,4-Benzoquinones	54	0.80	[28]
	1,4-Naphthoquinones	30	0.86	[28]
Topological indices	Aldehydes	6	n/a	[61]
	Benzenoids	23	0.97	[62]
	Indolizines	52	0.89	[63]
	Quinones	6	0.99	[61]

6.3 Optimal Descriptors in QSPR of Redox Potential

6.3.1 Basic Principles of Employing Optimal Descriptors in QSPR

Molecular descriptors derived from symbolic representations of the molecule are one of the very promising directions in QSPR because they do not contain statistical errors as experimentally derived descriptors [23]. The basic idea is to use molecular graphs to calculate descriptors that, being a representation of the molecular structure, can then be correlated with arbitrary physicochemical properties including the thermodynamic of the chemical species. Since their introduction in the 1980s, simplified molecular-input line-entry systems (SMILES) have represented an attractive alternative for the representation of the molecular structure by graph [64–66].

Currently, most molecule editors, computer programs for creating and modifying representations of chemical structures, support the conversion of graphical representations of chemical structure (i.e., topological information) to SMILES and vice versa. For the use of SMILES in QSPR, an efficient computer program CORAL [67] has been developed which is able to extract from the SMILES various graph theoretical invariants such as the vertex degree and the extended connectivity of higher order, as well as invariants for the graph of atomic orbitals. SMILES-based QSPR has been proven to be a powerful tool in the correlation of many physicochemical or biological properties [68].

A detailed description of CORAL and its use, including a discussion of the advantages and disadvantages of its use, is provided by Toropov et al. [69]. In a nutshell, the SMILES-based QSPR can be summarized as follows:

1. Collection of a set of chemical compounds and measurement of the desired physicochemical property (e.g., redox potential).
2. Conversion of the structure of the studied compounds into SMILES.
3. Calculation of the optimal descriptor of the correlation weight (DCW) as a mathematical function of SMILES, which is defined as,

$$\text{DCW} = \sum_{k=1}^N \text{CW}(S_k), \quad (6.13)$$

where S_k is a rule one-character fragment of the SMILES notation (situations where two symbols cannot be examined separately, e.g., 'Cl,' 'Br'), $\text{CW}(S_k)$ the so-called correlation weight of S_k , N is the number of characters in the given SMILES. The correlation weight $\text{CW}(S_k)$ is calculated by the Monte Carlo method [70] as coefficients which produce the largest correlation coefficient between the DCW and the endpoint examined of the training set. Using calculated $\text{CW}(S_k)$, it is possible to calculate DCW for training and test sets of all substances examined.

4. The QSPR model is then based on the least squares method,

$$(\text{Endpoint})_{\text{pred}} = C_0 + C_1 \times \text{DCW}, \quad (6.14)$$

where $(\text{Endpoint})_{\text{pred}}$ is the predictive endpoint, which can be validated with the structures of the test set, and C_0 and C_1 are regression coefficients.

6.3.2 *Published Studies on SMILES-Based QSPR for Redox Potential*

To date, only four QSPR studies using SMILES-based optimal descriptors have been published in the literature to correlate redox potential with the structure.

In 2006, the first SMILES-based QSPR analysis of half-wave potentials was performed for 40 benzoxazines, which belong to possible antituberculosic agents. Toropov et al. [71] have shown that the statistical quality of this approach ($R^2 = 0.882$) is fully comparable with the classical approach based on Hammett constants for the same data set ($R = 0.897$ [36]). This pilot study demonstrated and confirmed the suitability of using SMILES-based optimal descriptors for predicting the redox potentials of heterocyclic organic compounds.

In 2012, Toropov and Nesmerak [72] established SMILES-based QSPR for half-wave potential of 16 antimycobacterially active 1-phenyl-5-benzylsulfanyl-tetrazoles. The predictive potential of the applied approach was tested with three random splits into training and test sets, and $R^2 > 0.75$ was observed for all splits. The SMILES attributes, which are promoters of decrease of the half-wave potential in this QSPR, were identified.

This was followed in 2013 by a study by Nesmerak et al. [73] in which SMILES notation was used in QSPR of the half-wave potential of 24 derivatives of *N*-benzylsalicylthioamide. A detailed statistical evaluation of the predictive potential of the applied approach was carried out with three random splits into the sub-training, calibration, test, and validation sets. The $R^2 > 0.72$ was observed for all validation sets. Again, the SMILES attributes, which are promoters of an increase and decrease of the half-wave potential in this QSPR, were identified.

The most recent work published so far using SMILES-based optimal descriptors is the 2016 paper by Nesměřák et al. [38], which studied the half-wave potentials of 22 derivatives of 4-(benzylsulfanyl)pyridine. In the work, the QSPR approach using Hammett σ constants was compared with SMILES-based QSPR for three random distributions of derivatives into three sets (training, calibration, and validation). It was found that the SMILES-based equations have more validity from a statistical point of view (higher coefficients of determination); moreover, this approach allows one to identify the influence of individual structural motifs on the value of the half-wave potential.

6.3.3 Case Study of Two Large Data Sets

Here, the feasibility of using SMILES-based optimal descriptors in QSPR of the redox potential is demonstrated on two large data sets that have not been tested in this way before. Both data sets contain different chemical compounds with different numbers of atoms, which is reflected in the variability of their SMILES:

1. Data Set 1, which contains data on half-wave potentials for 71 aldehydes and ketones, has already been published by Garkani-Nejad and Rashidi-Nodeh [43]. In their study, the authors searched the QSPR for the half-wave potential using multiple linear regression, partial least square, artificial neural network, and wavelet neural network modeling methods. The best-established model was based on an artificial neural network and has $R^2 = 0.993$ for validation set.

Table 6.2 contains the data applied here to build up models using the Monte Carlo method.

2. Data Set 2 contains the data obtained in our previous work [38, 71–73] (which we mentioned in Sect. 6.3.2). The data applied here to build up models using the Monte Carlo method are tabulated in Table 6.3.

Data for each individual data set studied were haphazardly distributed into four sets: an active training set ($\approx 25\%$), a passive training set ($\approx 25\%$), a calibration set ($\approx 25\%$), and a validation set ($\approx 25\%$). The specific distribution of individual data to a given set is shown in Tables 6.2 and 6.3, respectively. The assignment of the active training set is to project the model. The molecular features, which are extracted from SMILES of this set, are included in the Monte Carlo optimization process to grant correlation weights that give the maximum correlation coefficient between the DCW and the half-wave potential. The passive training set is used to test whether the model projected from the active training set is acceptable for such SMILES that were not present in the active training set. The purpose of the calibration set is to detect the onset of the overtraining (overfitting). At the start of the optimization process, the correlation coefficients between the half-wave potential experimental values and DCW simultaneously increase for all sets, but the correlation coefficient for the calibration set attains a maximum; that is, the onset of overfitting is reached. The continuation of the optimization process results in a decrease of the correlation coefficient value for the calibration set. Thus, optimization procedure should be ceased when overtraining begins. After the Monte Carlo optimization procedure is completed, the validation set is employed to evaluate the predictive potential of the obtained model.

Models for both data sets were built using a single type of molecular descriptor, calculated as,

$$\text{DCW}(T, N) = \sum \text{CW}(S_k) + \sum \text{CW}(SS_k) + \sum \text{CW}(\text{EC}0_k) + \sum \text{CW}(\text{EC}1_k), \quad (6.15)$$

where the S_k is a SMILES-atom, i.e., single symbol in SMILES or a group of symbols which cannot be examined separately, and the SS_k is a pair of SMILES-atoms. The $\text{EC}0_k$ and $\text{EC}1_k$ are the Morgan extended connectivity of zero and first order, respectively. The $\text{CW}(x)$ is the correlation weights of the listed molecular features extracted from SMILES or the graph of atomic orbitals (GAO) [74]. Figure 6.1 and Table 6.4 contain an example of the adjacency matrix of GAO for compound #1 of Data Set 1, which is acetaldehyde.

The numerical data on the $\text{CW}(x)$ are calculated by the Monte Carlo method, which is the optimization process with the target function defined as,

$$\text{TF} = r_{\text{AT}} + r_{\text{PT}} - 0.1 |r_{\text{AT}} - r_{\text{PT}}| + 0.1 \text{IIC} + 0.5 \text{CII}, \quad (6.16)$$

where r_{AT} and r_{PT} are correlation coefficients between the observed and predicted endpoint for the active training set and the passive training set, respectively. The IIC

Table 6.2 Compounds in Data Set 1: ID, chemical name, CASRN, SMILES notation, experimental value of half-wave potential, spreading of the compound in a set (AT, PT, C, and V are active training, passive training, calibration, and validation sets, respectively), optimal descriptor of the correlation weight, predicted value of half-wave potential according to Eq. (6.17), difference between experimental and predicted value of half-wave potential, statistical SMILES-defect according to Eq. (6.20), and applicability of QSPR

ID	Name, CASRN	SMILES	$(-E_{1/2})_{\text{exp}}/V$	Set	DCW	$(-E_{1/2})_{\text{pred}}/V$	$\Delta((-E_{1/2})_{\text{exp}} - (-E_{1/2})_{\text{pred}})/V$	D	Applicability
1	Acetaldehyde, 75-07-0	O=CC	1.8900	AT	18.86460	1.6918	0.1982	2.4387	Yes
2	Benzaldehyde, 100-52-7	O=CC= C=CC=CC	1.3200	C	9.51991	1.3666	-0.0466	0.7209	Yes
3	Chloroacetaldehyde, 107-20-0	O=C(Cl)C	1.6600	PT	14.33960	1.5343	0.1257	0.2577	Yes
4	Crotonaldehyde, 4170-30-3	O=CC=CC	1.3000	V	15.62179	1.5790	-0.2790	1.4961	Yes
5	Dichloroacetaldehyde, 79-02-7	O=C(Cl)C(Cl)	1.6700	C	14.86018	1.5525	0.1175	0.2832	Yes
6	3,7-Dimethyl-2,6-octadienal, 5392-40-5	O=CC=C(C)CCC=C(C)C	2.2200	PT	19.28324	1.7064	0.5136	1.3337	Yes
7	2-Furaldehyde, 98-01-1	O=CC= OC=CC	1.4300	V	4.25814	1.1835	0.2465	0.7170	Yes
8	Glucose, 50-99-7	O=CC(O)C(O)C(O)C(O)CO	1.5500	V	16.66634	1.6153	-0.0653	2.3843	Yes
9	Glyceraldehyde, 56-82-6	O=CC(O)CO	1.5500	C	16.69810	1.6164	-0.0664	0.8468	Yes
10	Glycolaldehyde, 141-46-8	O=CCO	1.7000	PT	13.88409	1.5185	0.1815	0.4994	Yes
11	Glyoxal, 107-22-2	O=CC=O	1.4100	AT	18.28684	1.6717	-0.2617	0.3150	Yes
12	4-Hydroxy-2-methoxybenzaldehyde, 18278-34-7	O=C(Cl)=CC=C(O)C=C OC	1.4700	AT	15.89901	1.5886	-0.1186	1.6397	Yes
13	<i>o</i> -Methoxybenzaldehyde, 135-02-4	O=CC= C=CC=CC OC	1.4900	PT	2.99983	1.1397	0.3503	0.9444	Yes
14	2-Propenal, 107-02-8	O=CC=C	1.2900	C	8.29852	1.3241	-0.0341	0.3300	Yes
15	Phthalaldehyde, 643-79-8	O=CC= C=CC=CC C=O	1.3600	V	15.60209	1.5783	-0.2183	0.8390	Yes
16	Pyrrrole-2-carboxaldehyde, 1003-29-8	O=C(Cl)=CC=C N	1.2500	C	9.99636	1.3832	-0.1332	0.7717	Yes
17	Salicylaldehyde, 90-02-8	O=CC= C=CC=CC O	1.3200	V	3.30896	1.1505	0.1695	1.0718	Yes
18	Trichloroacetaldehyde, 75-87-6	O=C(Cl)C(Cl)Cl	1.6600	PT	14.43462	1.5377	0.1223	0.4853	Yes

(continued)

Table 6.2 (continued)

ID	Name, CASRN	SMILES	$(-E_{1/2})_{\text{exp}}/V$	Set	DCW	$(-E_{1/2})_{\text{pred}}/V$	$\Delta((-E_{1/2})_{\text{exp}} - (-E_{1/2})_{\text{pred}})/V$	D	Applicability
19	Bromoacetaldehyde, 17157-48-1	O=CBr	1.5800	AT	- 4.79514	0.8685	0.7115	19.3601	No
20	Acetone, 67-64-1	O=C(C)C	1.5200	AT	5.76596	1.2360	0.2840	0.4470	Yes
21	Acetophenone, 98-86-2	O=C(C=CC=CC)C	1.3300	V	7.94278	1.3117	0.0183	0.7928	Yes
22	Benzil, 134-81-6	O=C(C=CC=CC)C(=O)C=C	1.3600	C	7.69995	1.3033	0.0567	1.3753	Yes
23	Benzophenone, 119-61-9	O=C(C=CC=CC)C=CC=CC	0.5000	AT	2.94192	1.1377	- 0.6377	1.1315	Yes
24	Benzoin, 119-53-9	O=C(C=CC=CC)C(O)C=C	1.4900	AT	4.31901	1.1856	0.3044	1.5502	Yes
25	Bromoacetone, 598-31-2	O=C(C)CBr	0.2900	AT	- 10.22709	0.6794	- 0.3894	19.4650	No
26	3-Buten-2-one, 78-94-4	O=C(C=C)C	1.4200	PT	6.95839	1.2775	0.1425	0.3884	Yes
27	Butyphenone, 495-40-9	O=C(C=CC=CC)CCC	1.5500	C	18.41520	1.6762	- 0.1262	0.9164	Yes
28	Chloroacetone, 78-95-5	O=C(C)CCl	1.1800	C	8.93905	1.3464	- 0.1664	0.5128	Yes
29	Cyclohexanone, 108-94-1	O=C1CCCCC1	2.4500	PT	35.00197	2.2534	0.1966	0.4431	Yes
30	Anthrone, 90-44-8	O=C1C=CC=CC2C=CC=CC13	0.9300	V	- 3.65314	0.9082	0.0218	1.5757	Yes
31	1,5-Diphenyl-1,5-pentanedione, 6263-83-8	O=C(C=CC=CC)C(=O)C(=O)C=CC=CC	2.1000	AT	26.36662	1.9529	0.1471	1.4655	Yes
32	Flavanone, 487-26-3	O=C1C=CC=CC2OC(C=3C=CC=CC3)C1	1.4400	PT	5.28831	1.2194	0.2206	1.7246	Yes
33	Fluorescein, 2521-07-5	O=C1OC2(C3=CC=C(O)C=C3C=C4C=CC=CC4)C(=O)C=C15	1.5100	C	11.38388	1.4315	0.0785	3.6411	Yes
34	<i>o</i> -Hydroxyacetophenone, 118-93-4	O=C(C=CC=CC)OC	1.3600	V	1.79013	1.0976	0.2624	1.2516	Yes
35	<i>p</i> -Hydroxyacetophenone, 99-93-4	O=C(C=CC=C(O)C)C	1.4600	C	10.34086	1.3952	0.0648	1.5571	Yes
36	1,2,3-Indantrione, 938-24-9	O=C1C(=O)C=CC=CC2C1=O	1.3500	AT	3.22637	1.1476	0.2024	1.1989	Yes

(continued)

Table 6.2 (continued)

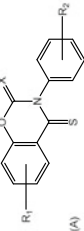
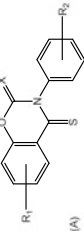
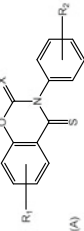
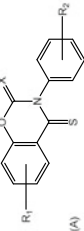
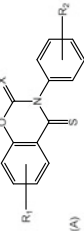
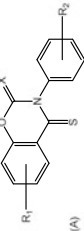
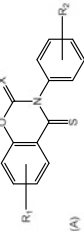
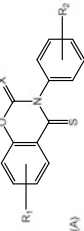
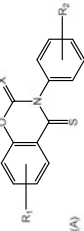
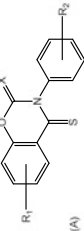
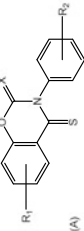
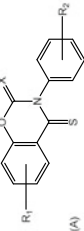
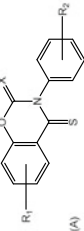
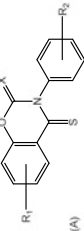
ID	Name, CASRN	SMILES	$(-E_{1/2})_{\text{exp}}/V$	Set	DCW	$(-E_{1/2})_{\text{pred}}/V$	$\Delta(((-E_{1/2})_{\text{exp}} - (-E_{1/2})_{\text{pred}})/V)$	D	Applicability
37	α -Iomone, 127-41-3	<chem>O=C(C=CC)C(=CCCC1(C)O)C</chem>	1.5900	V	12.04423	1.4545	0.1355	4.5733	No
38	Isatin, 91-56-5	<chem>O=C1NC=2C=CC=CC2C1=O</chem>	0.8000	PT	2.21803	1.1125	-0.3125	0.8255	Yes
39	4-Methyl-3,5-heptadien-2-one, 6263-81-6	<chem>O=C(C=C(C=CC)C</chem>	0.6400	AT	6.45439	1.2599	-0.6199	2.0861	Yes
40	4-Methyl-2,6-heptanedione, 5526-47-6	<chem>O=C(C)CC(C)CC(=O)C</chem>	1.2400	AT	10.94171	1.4161	-0.1761	1.3755	Yes
41	Pulegone, 89-82-7	<chem>O=C1C(=C(C)C)CC(C)C1</chem>	1.7400	AT	13.12445	1.4921	0.2479	1.5041	Yes
42	Testosterone, 58-22-0	<chem>O=C1C=C2CCC3C(C)CC4(C)C(O)CCC34)2(C)CC1</chem>	1.7900	AT	24.48299	1.8873	-0.0973	17.0565	No
43	1-Naphthalenecarboxaldehyde, 66-77-3	<chem>O=CC1=CC=CC=2C=CC=CC12</chem>	0.9100	V	-2.05872	0.9637	-0.0537	2.1434	Yes
44	2-Naphthalenecarboxaldehyde, 66-99-9	<chem>O=CC=1C=CC=2C=CC=CC2C1</chem>	0.9600	V	-0.15295	1.0300	-0.0700	1.2121	Yes
45	2-Phenanthrenecarboxaldehyde, 26842-00-2	<chem>O=CC=1C=CC2=C(C)C=CC=3C=CC=CC3)C1</chem>	1.0000	C	-4.80257	0.8682	0.1318	1.8349	Yes
46	3-Phenanthrenecarboxaldehyde, 7466-50-4	<chem>O=CC=1C=CC=2C=CC=3C=CC=CC3)C2C1</chem>	0.9400	C	-9.32433	0.7109	0.2291	1.7799	Yes
47	9-Phenanthrenecarboxaldehyde, 4707-71-5	<chem>O=CC1=CC=2C=CC=CC2C=3C=CC=CC13</chem>	0.8300	V	-5.85946	0.8314	-0.0014	1.5097	Yes
48	1-Anthracenecarboxaldehyde, 1140-79-0	<chem>O=CC1=CC=CC2=CC=3C=CC=CC3)C=C12</chem>	0.7500	AT	-2.69376	0.9416	-0.1916	2.8750	Yes
49	5,5-Dimethyl-3-phenyl-2-cyclohexen-1-one, 36047-17-3	<chem>O=C1C=C(C)C=2C=CC=CC2)CC(C)C1</chem>	1.7100	V	15.21679	1.5649	0.1451	1.3029	Yes
50	Cyclopentanone, 120-92-3	<chem>O=C1CCCC1</chem>	2.8200	PT	30.32197	2.0905	0.7295	0.3823	Yes
51	2,2-Dimethyl-1,3-diphenyl-1,3-propanedione, 41169-42-0	<chem>O=C(C)C(=C)C(=CC1)C(C)C(=O)C=2C=CC=CC2)C1C</chem>	1.8000	PT	14.85337	1.5522	0.2478	1.5869	Yes
52	Cinnamaldehyde, 104-55-2	<chem>O=CC=CC=1C=CC=CC1</chem>	1.7000	AT	15.86075	1.5873	0.1127	0.8775	Yes
53	Benzoylacetone, 93-91-4	<chem>O=C(C)C(=1C=CC=CC1)CC(=O)C</chem>	1.6800	AT	17.55251	1.6462	0.0338	1.1582	Yes

(continued)

Table 6.2 (continued)

ID	Name, CASRN	SMILES	$(-E_{1/2})_{\text{exp}}/V$	Set	DCW	$(-E_{1/2})_{\text{pred}}/V$	$\Delta(((-E_{1/2})_{\text{exp}} - (-E_{1/2})_{\text{pred}})/V)$	D	Applicability
54	2,3-Butanedione, 431-03-8	<chem>O=C(C(=O)C)C</chem>	0.8400	PT	9.05223	1.3503	-0.5103	0.6173	Yes
55	Coumarin, 91-64-5	<chem>O=C1OC=2C=CC=CC2C=C1</chem>	1.1100	V	-6.45597	0.8107	0.2993	0.9413	Yes
56	Dithizone, 60-10-6	<chem>S=C(N=NC=1C=CC=CC1)NNC=2C=CC=CC2</chem>	0.6000	PT	-3.31469	0.9200	-0.3200	3.3010	Yes
57	Fructose, 57-48-7	<chem>O=C(CO)C(O)C(O)C(O)CO</chem>	1.7600	AT	24.48262	1.8873	-0.1273	2.1525	Yes
58	4-Methyl-3-penten-2-one, 141-79-7	<chem>O=C(C=C(C)O)C</chem>	1.6000	AT	5.37453	1.2224	0.3776	0.7844	Yes
59	4-Phenyl-3-buten-2-one, 122-57-6	<chem>O=C(C=CC=1C=CC=CC1)C</chem>	1.2700	V	6.58019	1.2643	0.0057	1.1057	Yes
60	Phthalide, 87-41-2	<chem>O=C1OC(=O)C=CC=C1</chem>	0.2000	PT	-3.63589	0.9088	-0.7088	1.8389	Yes
61	2-Pyrenecarboxaldehyde, 26933-87-9	<chem>O=CC=1C=C2C=CC3=CC=CC=4C=CC(C1)=C2C34</chem>	1.0000	V	1.37032	1.0830	-0.0830	2.3661	Yes
62	Formaldehyde, 50-00-0	<chem>O=C</chem>	1.5900	PT	5.32980	1.2208	0.3692	0.0063	Yes
63	4-Hydroxybenzaldehyde, 123-08-0	<chem>O=CC1=CC=C(O)C=C1</chem>	1.4100	V	11.91799	1.4501	-0.0401	1.4851	Yes
64	<i>p</i> -Methoxybenzaldehyde, 123-11-5	<chem>O=CC1=CC=C(OC)C=C1</chem>	1.4800	C	11.12050	1.4223	0.0577	1.4193	Yes
65	Methylglyoxal, 78-98-8	<chem>O=CC(=O)C</chem>	0.8300	PT	10.41150	1.3976	-0.5676	0.5000	Yes
66	Propionaldehyde, 123-38-6	<chem>O=CCC</chem>	1.9300	PT	17.08032	1.6297	0.3003	0.3621	Yes
67	<i>d</i> -Carvone, 2244-16-8	<chem>O=C1C(=CCC(C(=C)O)C)C</chem>	1.7100	C	15.43042	1.5723	0.1377	1.4234	Yes
68	<i>trans</i> -dibenzoylethylene, 959-28-4	<chem>O=C1C(=CC(=O)C=C1C=CC=CC1)C=C2C=CC=CC2</chem>	1.5200	C	11.66667	1.4413	0.0787	1.4726	Yes
69	Phthalimide, 85-41-6	<chem>O=C1NC(=O)C=2C=CC=CC12</chem>	1.4000	PT	1.43337	1.0852	0.3148	1.8973	Yes
70	1-Pyrenecarboxaldehyde, 3029-19-4	<chem>O=CC1=CC=C2C=CC3=CC=CC=4C=CC1=C2C34</chem>	0.7600	C	-2.35703	0.9533	-0.1933	2.4213	Yes
71	1,3-Diphenyl-1,3-propanedione, 120-46-7	<chem>O=C(C(=1C=CC=CC1)CC(=O)C=2C=CC=CC2)</chem>	1.4200	C	13.36420	1.5004	-0.0804	1.3933	Yes

Table 6.3 Compounds in Data Set 2: ID, structure, SMILES notation, experimental value of half-wave potential, spreading of the compound in a set (AT, PT, C, and V are active training, passive training, calibration, and validation sets, respectively), optimal descriptor of the correlation weight, predicted value of half-wave potential according to Eq. (6.18), difference between experimental and predicted value of half-wave potential, statistical SMILES-defect according to Eq. (6.20), and applicability of QSPR

ID	Structure	R ₁	R ₂	X	SMILES	(-E _{1/2}) _{exp} /V	Set	DCW	(-E _{1/2}) _{pred} /V	$\Delta((-E_{1/2})_{\text{exp}} - (-E_{1/2})_{\text{pred}})/V$	D	Applicability
1		7-OCH ₃		O	<chem>COc1ccc2c(c1)OC(=O)N(C2=S)k3ccccc3</chem>	1.4200	AT	29.57213	1.3376	0.0824	0.6903	Yes
2		7-OCH ₃	4-F	O	<chem>Fe1cccc(cc1)N2C(=S)k3ccc(cc3OC2=O)OC</chem>	1.4300	PT	33.80638	1.4493	-0.0193	0.8024	Yes
3		7-OCH ₃	4-Br	O	<chem>Brc1cccc(cc1)N2C(=S)k3ccc(cc3OC2=O)OC</chem>	1.4400	AT	34.17889	1.4591	-0.0191	0.9552	Yes
4		7-OCH ₃	3-F	O	<chem>Fe1cccc(c1)N2C(=S)k3ccc(cc3OC2=O)OC</chem>	1.4450	PT	35.32005	1.4892	-0.0442	0.8215	Yes
5		7-OCH ₃	3-Cl	O	<chem>Clc1cccc(c1)N2C(=S)k3ccc(cc3OC2=O)OC</chem>	1.4500	AT	34.83947	1.4766	-0.0266	0.9119	Yes
6		7-CH ₃	4-CH ₃	O	<chem>Cc1ccc(cc1)N2C(=S)k3ccc(C)cc3OC2=O</chem>	1.4150	V	32.61086	1.4178	-0.0028	0.7094	Yes
7		6-CH ₃	4-CH ₃	O	<chem>Cc1ccc(cc1)N2C(=S)k3ccc(C)cc3OC2=O</chem>	1.4200	V	32.61086	1.4178	0.0022	0.7094	Yes
8		4-Br		O	<chem>Brc1cccc(cc1)N2C(=S)k3ccccc3OC2=O</chem>	1.4900	V	37.93747	1.5583	-0.0683	0.8371	Yes
9		6-OCH ₃	4-CH ₃	O	<chem>Cc1ccc(cc1)N2C(=S)k3ccc(cc3OC2=O)OC</chem>	1.4500	AT	31.77452	1.3957	0.0543	0.7485	Yes
10		6-OCH ₃	4-F	O	<chem>Fe1cccc(cc1)N2C(=S)k3ccc(cc3OC2=O)OC</chem>	1.4600	V	33.80638	1.4493	0.0107	0.8024	Yes
11		6-OCH ₃	4-Br	O	<chem>Brc1cccc(cc1)N2C(=S)k3ccc(cc3OC2=O)OC</chem>	1.4650	C	34.17889	1.4591	0.0059	0.9552	Yes
12		6-OCH ₃	4-Cl	O	<chem>Clc1cccc(cc1)N2C(=S)k3ccc(cc3OC2=O)OC</chem>	1.4700	C	34.11037	1.4573	0.0127	0.9240	Yes
13		6-OCH ₃	3-F	O	<chem>Fe1cccc(c1)N2C(=S)k3ccc(cc3OC2=O)OC</chem>	1.4800	C	35.32005	1.4892	-0.0092	0.8215	Yes
14		6-OCH ₃	4-CN	O	<chem>N#Cc1ccc(cc1)N2C(=S)k3ccc(cc3OC2=O)OC</chem>	1.5100	C	42.21905	1.6712	-0.1612	0.8872	Yes

(continued)

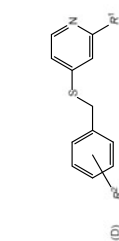
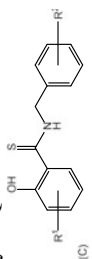


Table 6.3 (continued)

ID	Structure	R ₁	R ₂	X	SMILES	(-E _{1/2}) _{exp} /V	Set	DCW	(-E _{1/2}) _{pred} /V	$\Delta((-E_{1/2})_{\text{exp}} - (-E_{1/2})_{\text{pred}})/V$	D	Applicability
15	A	6-Cl		O	Clc1ccc2OC(=O)N(C(=S)c2c1)c3ccccc3	1.5300	AT	37.13971	1.5372	-0.0072	0.8072	Yes
16	A	6-Cl	3-Cl	O	Clc1cccc(c1)N2C(=S)c3ccc(Cl)ccc3OC2=O	1.5900	AT	38.92686	1.5844	0.0056	1.0625	Yes
17	A	7-OCH ₃	4-CH ₃	S	Cc1ccc(cc1)N2C(=S)c3ccc(cc3OC2=S)OC	1.2800	PT	28.57107	1.3112	-0.0312	0.6895	Yes
18	A	7-OCH ₃		S	COc1ccc2c(c1)OC(C(=S)N(C2=S)c3ccccc3	1.3150	V	26.36868	1.2531	0.0619	0.6313	Yes
19	A	7-OCH ₃	4-F	S	Fe1ccc(cc1)N2C(=S)c3ccc(cc3OC2=S)OC	1.3500	PT	30.60293	1.3648	-0.0148	0.7434	Yes
20	A	7-OCH ₃	4-Br	S	Brc1ccc(cc1)N2C(=S)c3ccc(cc3OC2=S)OC	1.3600	PT	30.97543	1.3746	-0.0146	0.8962	Yes
21	A	7-OCH ₃	4-Cl	S	Clc1ccc(cc1)N2C(=S)c3ccc(cc3OC2=S)OC	1.3700	C	30.90692	1.3728	-0.0028	0.8650	Yes
22	A	7-OCH ₃	3-F	S	Fe1ccc(cc1)N2C(=S)c3ccc(cc3OC2=S)OC	1.3900	V	32.11659	1.4047	-0.0147	0.7625	Yes
23	A	7-OCH ₃	3-Cl	S	Clc1ccc(cc1)N2C(=S)c3ccc(cc3OC2=S)OC	1.3950	AT	31.63602	1.3921	0.0029	0.8529	Yes
24	A	7-OCH ₃	4-CF ₃	S	FC(F)(F)c1ccc(cc1)N2C(=S)c3ccc(cc3OC2=S)OC	1.4050	C	33.58728	1.4435	-0.0385	1.3116	Yes
25	A	7-OCH ₃	3,4-Cl ₂	S	Clc1ccc(cc1)N2C(=S)c3ccc(cc3OC2=S)OC	1.4200	PT	33.46133	1.4402	-0.0202	8.0378	No
26	A	7-CH ₃	4-CH ₃	S	Cc1ccc(cc1)N2C(=S)c3ccc(C)ccc3OC2=S	1.3050	C	28.93306	1.3207	-0.0157	0.6522	Yes
27	A	6-CH ₃	4-CH ₃	S	Cc1ccc(cc1)N2C(=S)c3ccc(C)ccc3OC2=S	1.3200	V	28.93306	1.3207	-0.0007	0.6522	Yes
28	A		4-Br	S	Brc1ccc(cc1)N2C(=S)c3ccc3OC2=S	1.4200	C	34.25967	1.4613	-0.0413	0.7798	Yes
29	A	6-OCH ₃	4-CH ₃	S	Cc1ccc(cc1)N2C(=S)c3ccc(cc3OC2=S)OC	1.3300	V	28.57107	1.3112	0.0188	0.6895	Yes
30	A	6-OCH ₃		S	COc1ccc2OC(=S)N(C(=S)c2c1)c3ccccc3	1.3600	PT	28.09980	1.2988	0.0612	0.6422	Yes
31	A	6-OCH ₃	4-F	S	Fe1ccc(cc1)N2C(=S)c3ccc(cc3OC2=S)OC	1.3800	V	30.60293	1.3648	0.0152	0.7454	Yes
32	A	6-OCH ₃	4-Br	S	Brc1ccc(cc1)N2C(=S)c3ccc(cc3OC2=S)OC	1.4000	AT	30.97543	1.3746	0.0254	0.8962	Yes
33	A	6-OCH ₃	4-Cl	S	Clc1ccc(cc1)N2C(=S)c3ccc(cc3OC2=S)OC	1.4000	PT	30.90692	1.3728	0.0272	0.8650	Yes
34	A	6-OCH ₃	3-F	S	Fe1ccc(cc1)N2C(=S)c3ccc(cc3OC2=S)OC	1.4100	PT	32.11659	1.4047	0.0053	0.7625	Yes

(continued)

Table 6.3 (continued)

ID	Structure	R ₁	R ₂	X	SMILES	(-E _{1/2}) _{exp} /V	Set	DCW	(-E _{1/2}) _{pred} /V	$\Delta((E_{1/2})_{exp} - (-E_{1/2})_{pred})/V$	D	Applicability
35	A	6-OCH ₃	3-Cl	S	Clc1ccccc(c1)N2C(=S)c3ccc(ccc3OC2=S)OC	1.4300	AT	31.63602	1.3921	0.0379	0.8529	Yes
36	A	6-OCH ₃	4-CF ₃	S	FC(F)(F)c1ccc(cc1)N2C(=S)c3ccc(ccc3OC2=S)OC	1.4400	AT	33.58728	1.4435	-0.0035	1.3116	Yes
37	A	6-OCH ₃	3, 4-Cl ₂	S	Clc1ccccc(c1)N2C(=S)c3ccc(ccc3OC2=S)OC	1.4450	AT	33.46133	1.4402	0.0048	8.0378	No
38	A	6-OCH ₃	4-CN	S	N#Cc1ccc(cc1)N2C(=S)c3ccc(ccc3OC2=S)OC	1.4500	V	39.01560	1.5867	-0.1367	0.8282	Yes
39	A	6-Cl		S	Clc1ccc2OC(=S)N(C(=S)c2c1)c3ccccc3	1.4200	PT	33.93626	1.4527	-0.0327	0.7482	Yes
40	A	6-Cl	3-Cl	S	Clc1ccccc(c1)N2C(=S)c3ccc(Cl)ccc3OC2=S	1.5200	PT	35.24906	1.4874	0.0326	1.0052	Yes
41	B		4-OCH ₃	S	COc1ccc(cc1)CSc3nnnn3c2ccccc2	1.3700	PT	36.65209	1.5244	-0.1544	0.5216	Yes
42	B	4-Cl	4-OCH ₃	S	COc1ccc(cc1)CSc3nnnn3c2ccc(Cl)cc2	1.3900	AT	36.25180	1.5138	-0.1238	0.8023	Yes
43	B	4-Cl	3-OCH ₃	S	COc1ccc(cc1)CSc3nnnn3c2ccc(Cl)cc2	1.4300	V	37.76547	1.5538	-0.1238	0.8214	Yes
44	B	3, 4-Cl ₂	4-OCH ₃	S	COc1ccc(cc1)CSc3nnnn3c2cc(Cl)c(Cl)cc2	1.4000	AT	34.83135	1.4763	-0.0763	7.0087	No
45	B	4-CH ₃	4-OCH ₃	S	COc1ccc(cc1)CSc3nnnn3c2ccc(C)cc2	1.4000	C	32.21618	1.4074	-0.0074	0.5815	Yes
46	B	4-OCH ₃	4-Cl	S	Clc1ccc(cc1)CSc3nnnn3c2ccc(OC)cc2	1.5800	AT	37.61106	1.5497	0.0303	0.7557	Yes
47	B	4-OCH ₃	4-CF ₃	S	FC(F)(F)c1ccc(cc1)CSc3nnnn3c2ccc(OC)cc2	1.5900	C	40.29141	1.6204	-0.0304	1.2023	Yes
48	B	4-OCH ₃	4-OCH ₃	S	COc1ccc(cc1)CSc3nnnn3c2ccc(OC)cc2	1.3750	C	30.99003	1.3750	-0.0000	0.6186	Yes
49	B	4-OCH ₃		S	COc1ccc(cc1)nc3nnnc3Cc2ccccc2	1.5650	PT	38.53034	1.5739	-0.0089	0.5180	Yes
50	B	4-OCH ₃	4-CH ₃	S	Cc1ccc(cc1)CSc3nnnn3c2ccc(OC)cc2	1.5500	PT	35.27521	1.4881	0.0619	0.5802	Yes
51	B	4-Br	4-OCH ₃	S	COc1ccc(cc1)CSc3nnnn3c2ccc(Br)cc2	1.4050	AT	35.10903	1.4837	-0.0787	0.8466	Yes
52	B	4-OCH ₃	4-F	S	Fe1ccc(cc1)CSc3nnnn3c2ccc(OC)cc2	1.5800	AT	37.30706	1.5417	0.0388	0.6341	Yes
53	B	2-OCH ₃	4-OCH ₃	S	COc1ccc(cc1)CSc3nnnn3c2ccccc2OC	1.6300	C	37.77871	1.5541	0.0759	0.5996	Yes
54	B	2-OCH ₃	4-Cl	S	Clc1ccc(cc1)CSc3nnnn3c2ccccc2O	1.7800	V	43.24671	1.6984	0.0816	0.7785	Yes

(continued)

Table 6.3 (continued)

ID	Structure	R ₁	R ₂	X	SMILES	(-E _{1/2}) _{exp} /V	Set	DCW	(-E _{1/2}) _{pred} /V	$\Delta(((-E_{1/2})_{\text{exp}} - (-E_{1/2})_{\text{pred}})/V)$	D	Applicability
55	B	2-OCH ₃			Oc1ccccln3mnm3SCc2cccce2	1.7600	C	43.03163	1.6927	0.0673	0.5252	Yes
56	B	2-OCH ₃	3-Cl		Clc1ccccl(c1)CSs3mnm3c2cccce2O	1.8000	C	43.97580	1.7176	0.0824	0.7664	Yes
57	B	2-OCH ₃	4-F		Fe1ccc(cc1)CSs3mnm3c2cccce2O	1.8000	PT	42.94271	1.6903	0.1097	0.6569	Yes
58	B	2-OCH ₃	4-Br		Brc1ccc(cc1)CSs3mnm3c2cccce2O	1.8150	PT	43.31522	1.7002	0.1148	0.8097	Yes
59	B	2-OCH ₃	4-CH ₃		Cc1ccc(cc1)CSs3mnm3c2cccce2O	1.7600	V	40.91086	1.6367	0.1233	0.6030	Yes
60	C				S=C(NC)c1ccccl1)c2cccce2O	1.1670	C	26.91222	1.2674	-0.1004	0.5157	Yes
61	C		3-CH ₃		S=C(NC)c1ccc(C)c1)c2cccce2O	1.1650	PT	23.98997	1.1903	-0.0253	0.5947	Yes
62	C		3-Cl		Clc2ccc(CNC(=S)c1cccclO)ccc2	1.1900	V	26.55914	1.2581	-0.0681	0.7252	Yes
63	C		4-CH ₃		S=C(NC)c1ccc(C)cc1)c2cccce2O	1.1690	AT	22.47631	1.1504	0.0186	0.5756	Yes
64	C		4-Cl		Clc2ccc(CNC(=S)c1cccclO)ccc2	1.1540	C	25.83004	1.2389	-0.0849	0.7373	Yes
65	C		4-F		S=C(NC)c1ccc(F)cc1)c2cccce2O	1.1870	C	24.06275	1.1923	-0.0053	0.6659	Yes
66	C		4-OCH ₃		S=C(NC)c1ccc(OC)cc1)c2cccce2O	1.1170	C	21.25016	1.1181	-0.0011	0.6127	Yes
67	C		4- <i>tert</i> -butyl		S=C(NC)c1ccc(cc1)C(C)(C)c2cccce2O	1.1750	C	24.51218	1.2041	-0.0291	0.7955	Yes
68	C		3,4-Cl ₂		Clc2ccc(CNC(=S)c1cccclO)ccc2Cl	1.2030	V	26.22853	1.2494	-0.0464	6.8839	No
69	C	4-OCH ₃			S=C(NC)c1ccccl1)c2ccc(cc2O)OC	1.1320	PT	23.15364	1.1683	-0.0363	0.6338	Yes
70	C	5-Cl			Oc2ccc(Cl)ccc2C(=S)NCc1ccccl	1.2160	V	25.90747	1.2409	-0.0249	0.7761	Yes
71	C	4-CH ₃			S=C(NC)c1ccccl1)c2ccc(C)cc2O	1.1370	AT	23.98997	1.1903	-0.0533	0.5947	Yes
72	C	5-Br	3-Br		Oc2ccc(Br)ccc2C(=S)NCc1ccccl(Br)c1	1.2200	PT	25.57068	1.2320	-0.0120	1.1680	Yes
73	C	4-OCH ₃	3-Cl		Clc2ccc(CNC(=S)c1cccclO)ccc2	1.1250	PT	22.41074	1.1487	-0.0237	0.8414	Yes
74	C	4-CH ₃	3-NO ₂		S=C(NC)c1ccccl(c1)N+([O-])=O)c2ccc(C)cc2O	1.1900	PT	24.52111	1.2044	-0.0144	1.3409	Yes

(continued)

Table 6.3 (continued)

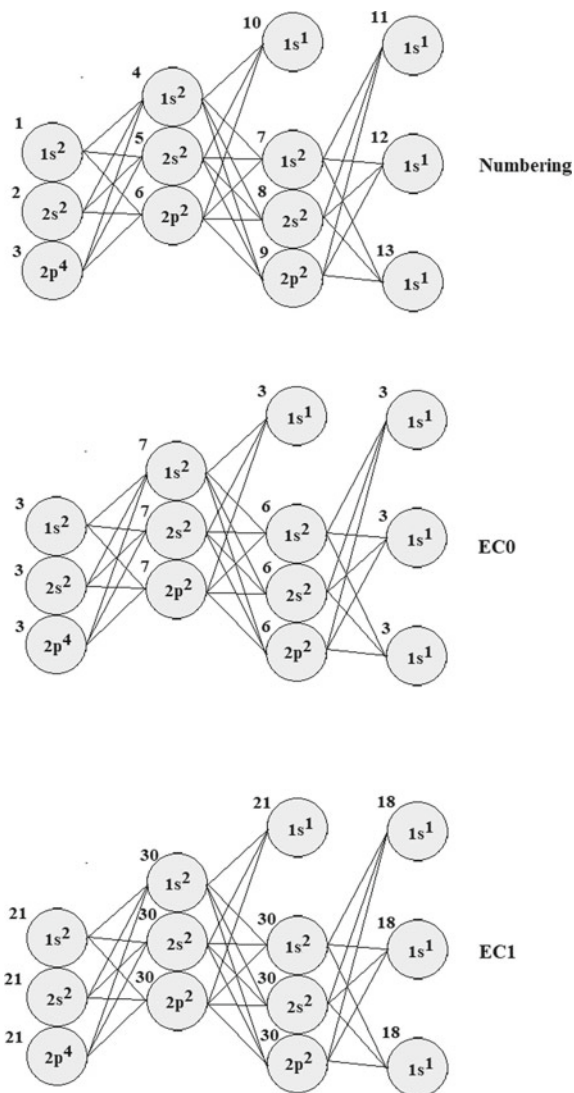
ID	Structure	R ₁	R ₂	X	SMILES	(-E _{1/2}) _{exp} /V	Set	DCW	(-E _{1/2}) _{pred} /V	$\Delta(((-E_{1/2})_{exp} - (-E_{1/2})_{pred})/V)$	D	Applicability
75	C	4-Cl	4-Br		BrC2ccc(CNC(=S)Cl)ccc(Cl)cc1O)cc2	1.2050	C	26.22737	1.2494	-0.0444	1.0372	Yes
76	C	5-Br	4-Br		Oc2ccc(Br)cc2C(=S)NCc1ccc(Br)cc1	1.1920	V	24.03161	1.1914	0.0006	1.1627	Yes
77	C	3-CH ₃	4-Cl		Clc2ccc(CNC(=S)Cl)ccc(C)clO)cc2	1.2300	C	25.17760	1.2217	0.0083	0.7890	Yes
78	C	5-Cl	4-F		Oc2ccc(Cl)cc2C(=S)NCc1ccc(F)cc1	1.2350	V	23.05800	1.1658	0.0692	0.9264	Yes
79	C	4-CH ₃	4-CH ₃		S=C(NCc1ccc(C)cc1)c2ccc(C)cc2O	1.1250	V	19.55406	1.0733	0.0517	0.6547	Yes
80	C	5-NO ₂	4-CH ₃		S=C(NCc1ccc(C)cc1)c2ccc(ccc2O)[N+](=O)=O	1.2310	V	23.00744	1.1644	0.0666	1.3218	Yes
81	C	5-Cl	3,4-Cl ₂		Clc2ccc(CNC(=S)Cl)ccc(Cl)ccc1O)cc2Cl	1.2440	PT	26.55734	1.2581	-0.0141	7.1526	No
82	C	5-Br	3,4-Cl ₂		Oc2ccc(Br)cc2C(=S)NCc1ccc(Cl)cc1	1.2180	AT	23.75393	1.1841	0.0339	7.3249	No
83	C	4-CH ₃	4- <i>tert</i> -butyl		S=C(NCc1ccc(cc1)C(C)(O)c2ccc(C)cc2O	1.1550	PT	21.58993	1.1270	0.0280	0.8745	Yes
84	D	-CN	-CN		N#Cc2cc(SCc1cccc1)ccn2	1.7200	AT	45.55175	1.7592	-0.0392	0.5422	Yes
85	D	-CN	3-Cl		N#Cc2cc(SCc1cccc1)ccn2	1.7700	V	45.88056	1.7678	0.0022	0.8108	Yes
86	D	-CN	4-Cl		N#Cc2cc(SCc1cccc(Cl)ccn2	1.7300	AT	45.15147	1.7486	-0.0186	0.8229	Yes
87	D	-CN	3-F		N#Cc2cc(SCc1cccc(F)c1)ccn2	1.7550	C	44.21595	1.7239	0.0311	0.7115	Yes
88	D	-CN	4-F		N#Cc2cc(SCc1cccc(F)cc1)ccn2	1.7250	AT	42.70228	1.6840	0.0410	0.6924	Yes
89	D	-CN	3-Br		N#Cc2cc(SCc1cccc(Br)cc1)ccn2	1.7700	V	45.54776	1.7591	0.0109	0.8724	Yes
90	D	-CN	4-Br		N#Cc2cc(SCc1cccc(Br)cc1)ccn2	1.7400	C	44.00869	1.7185	0.0215	0.8672	Yes
91	D	-CN	3-CH ₃		N#Cc2cc(SCc1cccc(C)c1)ccn2	1.7000	C	42.62951	1.6821	0.0179	0.6212	Yes
92	D	-CN	3-NO ₂		N#Cc2cc(SCc1cccc(c1)[N+](=O)=O)ccn2	1.7900	AT	46.08289	1.7732	0.0168	1.2883	Yes
93	D	-CN	4-NO ₂		N#Cc2cc(SCc1cccc(c1)[N+](=O)=O)ccn2	1.8100	AT	44.56922	1.7332	0.0768	1.2692	Yes
94	D	-CN	3,4-Cl ₂		N#Cc2cc(SCc1cccc(Cl)cc1)ccn2	1.7700	PT	43.73101	1.7111	0.0589	7.0293	No

(continued)

Table 6.3 (continued)

ID	Structure	R ₁	R ₂	X	SMILES	(-E _{1/2}) _{exp} /V	Set	DCW	(-E _{1/2}) _{pred} /V	$\Delta(((-E_{1/2})_{\text{exp}} - (-E_{1/2})_{\text{pred}})/V)$	D	Applicability
95	D	- CN	3,4-F ₂		N#C=C2cc(SCc1ccc(F)c1)cen2	1.7600	C	44.39381	1.7286	0.0314	0.9000	Yes
96	D	- CN	3,5-(NO ₂) ₂		N#C=C2cc(SCc1cc(cc(c1)N+)(O-)=O)N+](O-)=O)cen2	1.8500	PT	48.12769	1.8271	0.0229	2.0536	Yes
97	D	- CSNH ₂	3-Cl		S=C(N)c2cc(SCc1ccc(Cl)c1)cen2	1.1950	C	25.92540	1.2414	-0.0464	0.9052	Yes
98	D	- CSNH ₂	3-F		S=C(N)c2cc(SCc1ccc(F)c1)cen2	1.1850	C	24.26079	1.1975	-0.0125	0.8059	Yes
99	D	- CSNH ₂	3-Br		S=C(N)c2cc(SCc1ccc(Br)c1)cen2	1.2100	AT	25.59260	1.2326	-0.0226	0.9668	Yes
100	D	- CSNH ₂	3-NO ₂		S=C(N)c2cc(SCc1ccc(c1)N+)(O-)=O)cen2	1.2150	V	26.12773	1.2467	-0.0317	1.3827	Yes
101	D	- CSNH ₂	4-NO ₂		S=C(N)c2cc(SCc1ccc(cc1)N+)(O-)=O)cen2	1.2000	C	24.61406	1.2068	-0.0068	1.3636	Yes
102	D	- CSNH ₂	4-OCH ₃		S=C(N)c2cc(SCc1ccc(OC)c1)cen2	1.1300	V	19.93453	1.0834	0.0466	0.7336	Yes
103	D	- CSNH ₂	3,4-Cl ₂		S=C(N)c2cc(SCc1ccc(Cl)c(Cl)c1)cen2	1.1850	V	23.77586	1.1847	0.0003	7.1237	No
104	D	- CSNH ₂	3,4-F ₂		S=C(N)c2cc(SCc1ccc(F)c(F)c1)cen2	1.2200	PT	24.43865	1.2022	0.0178	0.9944	Yes
105	D	- CSNH ₂	3,5-(NO ₂) ₂		S=C(N)c2cc(SCc1cc(cc(c1)N+)(O-)=O)N+)(O-)=O)cen2	1.2300	PT	28.17253	1.3007	-0.0707	2.1479	Yes

Fig. 6.1 An example of a graph of atomic orbitals for compound #1 of Data Set 1 (acetaldehyde, SMILES is O=CC)



is the index of ideality of correlation [75], and the CII is the correlation intensity index [76].

QSPRs based on hybrid optimal descriptors were performed for both data sets examined. Table 6.5 contains an example of calculation of the DCW(1, 15) for compound #1 of Data Set 1 (acetaldehyde, SMILES is O=CC). The definition of DCW(1, 15) is the follows: (i) the threshold to define minimal number of the molecular features extracted from SMILES or from GAO in the training set (this is 1)

Table 6.4 Adjacency matrix of the graph of atomic orbitals for compound #1 of Data Set 1 (acetaldehyde, SMILES is O=CC)

		1	2	3	4	5	6	7	8	9	10	11	12	13	ECO	EC1
		1s ²	2s ²	2p ⁴	1s ²	2s ²	2p ²	1s ²	2s ²	2p ²	1s ¹	1s ¹	1s ¹	1s ¹		
1	1s ²	0	0	0	1	1	1	0	0	0	0	0	0	0	3	21
2	2s ²	0	0	0	1	1	1	0	0	0	0	0	0	0	3	21
3	2p ⁴	0	0	0	1	1	1	0	0	0	0	0	0	0	3	21
4	1s ²	1	1	1	0	0	0	1	1	1	1	0	0	0	7	30
5	2s ²	1	1	1	0	0	0	1	1	1	1	0	0	0	7	30
6	2p ²	1	1	1	0	0	0	1	1	1	1	0	0	0	7	30
7	1s ²	0	0	0	1	1	1	0	0	0	0	1	1	1	6	30
8	2s ²	0	0	0	1	1	1	0	0	0	0	1	1	1	6	30
9	2p ²	0	0	0	1	1	1	0	0	0	0	1	1	1	6	30
10	1s ¹	0	0	0	1	1	1	0	0	0	0	0	0	0	3	21
11	1s ¹	0	0	0	0	0	0	1	1	1	0	0	0	0	3	18
12	1s ¹	0	0	0	0	0	0	1	1	1	0	0	0	0	3	18
13	1s ¹	0	0	0	0	0	0	1	1	1	0	0	0	0	3	18

and (ii) the number of iterations in the Monte Carlo optimization for the correlation weights (this is 15).

The following QSPR equations were obtained:

1. for Data Set 1

$$-(E_{1/2})_{\text{pred}} = 1.0353363(\pm 0.0412942) + 0.0347993(\pm 0.0023327) \times \text{DCW}(1, 15) \quad (6.17)$$

2. for Data Set 2

$$-(E_{1/2})_{\text{pred}} = 0.5574662(\pm 0.0099109) + 0.0263809(\pm 0.0002835) \times \text{DCW}(1, 15) \quad (6.18)$$

Table 6.6 contains the statistical quality of these models. Figure 6.2 shows a graphical comparison of the correlation between the experimental and predicted values of the half-wave potential for the validation set for both data sets studied. The model for Data Set 1, published by Garkani-Nejad and Rashidi-Nodeh [43], is statistically better, but the model which is given by Eq. (6.17) is based on representation of the molecular structure solely by SMILES (GAO is extracted from SMILES using CORAL software).

The scope of applicability of the CORAL model is defined by the so-called statistical defects of the SMILES attributes [77]. These defects are calculated as,

Table 6.5 Calculation of the DCW(1, 15) for compound #1 of Data Set 1 (acetaldehyde, SMILES is O=CC)

x	CW(x)
EC0-1s ² 3...	- 0.6596
EC0-2s ² 3...	0.0746
EC0-2p ⁴ 3...	2.2005
EC0-1s ² 7...	0.4244
EC0-2s ² 7...	- 0.1800
EC0-2p ² 7...	- 0.0190
EC0-1s ² 6...	- 0.4280
EC0-2s ² 6...	- 0.7785
EC0-2p ² 6...	0.1469
EC0-1s ¹ 3...	0.8279
EC0-1s ¹ 3...	0.8279
EC0-1s ¹ 3...	0.8279
EC0-1s ¹ 3...	0.8279
EC1-1s ² 21...	1.6104
EC1-2s ² 21...	1.9570
EC1-2p ⁴ 21...	0.7659
EC1-1s ² 30...	1.9697
EC1-2s ² 30...	1.7082
EC1-2p ² 30...	1.2908
EC1-1s ² 30...	1.9697
EC1-2s ² 30...	1.7082
EC1-2p ² 30...	1.2908
EC1-1s ¹ 21...	- 0.4715
EC1-1s ¹ 18...	- 0.2068
EC1-1s ¹ 18...	- 0.2068
EC1-1s ¹ 18...	- 0.2068
O.....	1.6307
=.....	- 0.4282
C.....	- 0.1585
C.....	- 0.1585
O...=.....	1.0734
C...=.....	- 0.0588
C...C.....	- 0.3071
DCW(1, 15)	18.865

Table 6.6 Statistical characteristics of the models for the data sets examined for (AT) active training set, (PT) passive training set, (C) calibration set, and (V) validation set

		n	R^2	CCC	IIC	CII	Q^2	RMSE	MAE	F
Data Set 1	AT	19	0.5257	0.6891	0.6525	0.7187	0.3882	0.343	0.276	19
	PT	18	0.7412	0.7171	0.5221	0.8367	0.6584	0.403	0.346	46
	C	17	0.7952	0.8914	0.8917	0.9442	0.6983	0.122	0.106	58
	V	17	0.6679					0.165	0.124	
Data Set 2	AT	25	0.9301	0.9638	0.7578	0.9499	0.9191	0.049	0.038	306
	PT	27	0.9431	0.9635	0.6634	0.9619	0.9335	0.054	0.040	415
	C	28	0.9543	0.9702	0.9761	0.9684	0.9458	0.052	0.035	543
	V	25	0.9153					0.061	0.043	

Abbreviations used: n is the number of compounds in the corresponding set, R^2 is determination coefficient, CCC is the concordance correlation coefficient, IIC is the index of ideality of the correlation, CII is the correlation intensity index, Q^2 is cross-validated R^2 , RMSE is root mean squared error, MAE is the mean absolute error, and F is the Fischer F -ratio

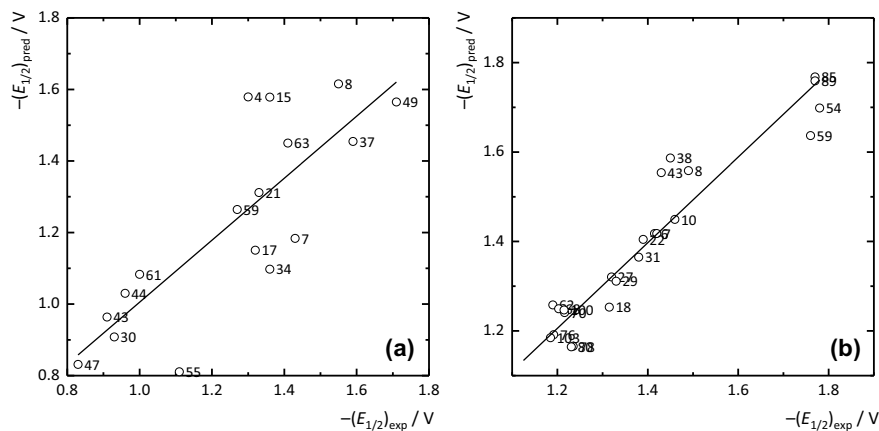


Fig. 6.2 Correlation between the experimental and predicted values of the half-wave potential for the validation set of **a** Data Set 1 and **b** Data Set 2. The numbers at each point indicate the compound numbers in Table 6.2, resp. Table 6.3. On the abscissa are plotted the experimentally measured values of the half-wave potential, on the ordinate are the values calculated according to Eqs. (6.17) and (6.18), respectively

$$d_k = \frac{|P(A_k) - P'(A_k)|}{N(A_k) - N'(A_k)} + \frac{|P(A_k) - P''(A_k)|}{N(A_k) - N''(A_k)} + \frac{|P'(A_k) - P''(A_k)|}{N'(A_k) - N''(A_k)}, \quad (6.19)$$

where $P(A_k)$, $P'(A_k)$, and $P''(A_k)$ are the probability of A_k in the active training set, the passive training set, and the calibration set, respectively; $N(A_k)$, $N'(A_k)$, and $N''(A_k)$ are the frequencies of A_k in the active training set, the passive training set, and the calibration set, respectively. The statistical SMILES-defects (D_j) are calculated as

$$D_j = \sum_{k=1}^{\text{NA}} d_k, \quad (6.20)$$

where NA is the number of non-blocked SMILES attributes in the SMILES. A given SMILES falls in the domain of applicability if

$$D_j < 2\bar{D} \quad (6.21)$$

The \bar{D} is average value of the statistical defect on the association of the active training set, passive training set, and calibration set. As can be seen from Table 6.2, the model for Data Set 1, given by Eq. (6.17), cannot be applied to compounds #19, #25, #37, and #42, which represent 5.6% of the data set. In the case of Data Set 2, the model given by Eq. (6.18) is not applicable for compounds #25, #37, #44, #68, #81, #82, #94, and #103; that is 7.6% in total (Table 6.3). For this model, it is interesting that the compounds that are excluded are all substituted on one of the ring moieties by two chlorine atoms; the probable reason is that chlorine is a substituent with a very strong negative induction effect.

From the statistical quality of the QSPRs obtained, it can be seen that the model derived for Data Set 2 has a higher statistical validity. This is probably due to the smaller variability in the SMILES of the individual compounds included in Data Set 2 compared to Data Set 1. In Data Set 1, the length of the SMILES ranges from 3 to 51 characters while in Data Set 2 the length ranges from 23 to 54 characters.

Finding successful QSPR models between the half-wave potential and the structure of the molecule for both data sets studied demonstrated that optimal descriptors calculated with molecular features extracted from SMILES together with molecular features extracted from GAO are very useful molecular descriptors applicable to QSPR of non-congeneric and structurally diverse compounds (which is a very topical issue in QSAR/QSPR [78]).

6.4 Conclusions

As the redox potential is an important electrochemical property used for the characterization of chemical species, this chapter illustrates the possibilities of using hybrid optimal descriptors calculated with molecular features extracted from SMILES together with molecular features extracted from GAO for developing QSPR models for redox potential. The CORAL software is able to be an efficient tool for building a robust model for redox potentials of various classes of compounds. On two large data sets, it was found that although the sets contained structurally very different substances, statistically significant correlations could be found. The quality of the correlations is affected by the difference in the number of features that form SMILES. It has also been confirmed that an optimal descriptor can be a translator of eclectic information into a model for the prediction of redox potential.

References

1. Marcus RA (1993) *Rev Mod Phys* 65:599–610. <https://doi.org/10.1103/RevModPhys.65.599>
2. Bard AJ, Faulkner LR White HS (2022) *Electrochemical methods: fundamentals and applications*, 3rd edn. Wiley, New York
3. Amini A, Harriman A (2003) *J Photochem Photobiol C* 4:155–177. [https://doi.org/10.1016/S1389-5567\(03\)00027-3](https://doi.org/10.1016/S1389-5567(03)00027-3)
4. Lister SG, Reynolds CA, Richards WG (1992) *Int J Quantum Chem* 41:293–310. <https://doi.org/10.1002/qua.560410206>
5. Mauk AG (1999) *Essays Biochem* 34:101–124. <https://doi.org/10.1042/bse0340101>
6. Nesměrāk K (2020) *Mini-Rev Med Chem* 20:1341–1356. <https://doi.org/10.2174/1389557520666200204121806>
7. Francke R, Little RD (2014) *Chem Soc Rev* 43:2492–2521. <https://doi.org/10.1039/c3cs60464k>
8. Ottosen LM, Larsen TH, Jensen PE, Kirkelund GM, Kerm-Jespersen H, Tuxen N, Hyldegaard BH (2019) *Chemosphere* 235:113–125. <https://doi.org/10.1016/j.chemosphere.2019.06.075>
9. Montcourrier P, Silver I, Farnoud R, Bird I, Rochefort H (1997) *Clin Exp Metastasis* 15:382–392. <https://doi.org/10.1023/A:1018446104071>
10. Klibanov AM (2001) *Nature* 409:241–246. <https://doi.org/10.1038/35051719>
11. Doukyu N, Ogino H (2010) *Biochem Eng J* 48:270–282. <https://doi.org/10.1016/j.bej.2009.09.009>
12. Hillard EA, de Abreu FC, Ferreira DCM, Jaouen G, Goulart MOF, Amatore C (2008) *Chem Commun (Camb)* 2612–2628. <https://doi.org/10.1039/b718116g>
13. Hammerich O, Speiser B (2016) *Organic electrochemistry*, 5th edn. CRC Press, Boca Raton
14. Appleby AJ, Zagal JH (2011) *J Solid State Electrochem* 15:1811–1832. <https://doi.org/10.1007/s10008-011-1394-8>
15. Hansch C, Leo A (1995) *Exploring QSAR: fundamentals and applications in chemistry and biology*. American Chemical Society, Washington
16. Carloni P, Alber F (2003) *Quantum medicinal chemistry*. Wiley, Weinheim. <https://doi.org/10.1002/3527602712>
17. Mercader AG, Duchowicz PR, Sivakumar P (2016) *Chemometrics applications and research: QSAR in medicinal chemistry*. Apple Academic Press, Oakville
18. Williams A (2003) *Free energy relationships in organic and bio-organic chemistry*. Royal Society of Chemistry, London
19. Roy K, Kar S, Das RN (2015) *A primer on QSAR/QSPR modeling: fundamental concepts*. Springer, Cham. <https://doi.org/10.1007/978-3-319-17281-1>
20. Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, Öberg T, Todeschini R, Fourches D, Varnek A (2008) *J Chem Inf Model* 48:1733–1746. <https://doi.org/10.1021/ci800151m>
21. Zhao L, Wang W, Sedykh A, Zhu H (2017) *ACS Omega* 2:2805–2812. <https://doi.org/10.1021/acsomega.7b00274>
22. Young D, Martin T, Venkatapathy R, Harten P (2008) *QSAR Comb Sci* 27:1337–1345. <https://doi.org/10.1002/qsar.200810084>
23. Todeschini R, Consonni V (2009) *Molecular descriptors for chemoinformatics*, vols 1 and 2. Wiley-VCH, Weinheim. <https://doi.org/10.1002/9783527628766>
24. Heyrovsky J (1934) *A polarographic study of the electro-kinetic phenomena of adsorption, electro-reduction and overpotential displayed at the dropping mercury cathode*. Hermann & Cie, Paris
25. Shikata M, Tachi I (1938) *Collect Czech Chem Commun* 10:368–379. <https://doi.org/10.1135/cccc19380368>
26. Hammett LP (1938) *Trans Faraday Soc* 34:156–165. <https://doi.org/10.1039/TF9383400156>
27. Zuman P (1967) *Substituent effects in organic polarography*. Plenum Press, New York. <https://doi.org/10.1007/978-1-4684-8661-2>

28. Driebergen RJ, Moret EE, Janssen LHM, Blauw JS, Holthus JMM, Postma Kelder SJ, Verboom W, Reinhoudt DN, van der Linden WE (1992) *Anal Chim Acta* 257:257–273. [https://doi.org/10.1016/0003-2670\(92\)85179-A](https://doi.org/10.1016/0003-2670(92)85179-A)
29. Elhabiri M, Sidorov P, Cesar-Rodo E, Marcou G, Lanfranchi DA, Davioud-Charvet E, Horvath D, Varnek A (2015) *Chem Eur J* 21:3415–3424. <https://doi.org/10.1002/chem.201403703>
30. Beheshti A, Norouzi P, Ganjali MR (2012) *Int J Electrochem Sci* 7:4811–4821
31. Goudarzi N, Goodarzi M, Hosseini MM, Nekooei M (2009) *Mol Phys* 107:1739–1744. <https://doi.org/10.1080/00268970903042266>
32. Mohammadhossein M, Nekoei M (2013) *Asian J Chem* 25:349–352. <https://doi.org/10.14233/ajchem.2013.13061>
33. Noorizadeh H, Farmany A (2015) *Russ J Electrochem* 51:249–257. <https://doi.org/10.1134/S102319351503009X>
34. Hemmateenejad B, Yazdani M (2009) *Anal Chim Acta* 634:27–35. <https://doi.org/10.1016/j.aca.2008.11.062>
35. Jurlina JL, Lindsay A, Packer JE, Baguley BC, Denny WA (1987) *J Med Chem* 30:473–480. <https://doi.org/10.1021/jm00386a006>
36. Nesmerak K, Nemeč I, Sticha M, Waissner K, Palat K (2005) *Electrochim Acta* 50:1431–1437. <https://doi.org/10.1016/j.electacta.2004.08.031>
37. Wang LY, Cao CT, Cao CZ (2016) *Chin J Chem Phys* 29:260–264. <https://doi.org/10.1063/1674-0068/29/cjcp1508173>
38. Nesměrák K, Toropov AA, Toropova AP (2016) *J Electroanal Chem* 766:24–29. <https://doi.org/10.1016/j.jelechem.2016.01.032>
39. Pelmus M, Ungureanu EM, Stanescu MD, Tarko L (2020) *J Appl Electrochem* 50:851–862. <https://doi.org/10.1007/s10800-020-01417-0>
40. Tömpe P, Clementis G, Petneházy I, Jászay ZM, Töke L (1995) *Anal Chim Acta* 305:295–303. [https://doi.org/10.1016/0003-2670\(94\)00354-O](https://doi.org/10.1016/0003-2670(94)00354-O)
41. Moraleda D, El Abed D, Pellissier H, Santelli M (2006) *J Mol Struct THEOCHEM* 760:113–119. <https://doi.org/10.1016/j.theochem.2005.12.001>
42. Hadjmohammadi MR, Kamel K, Biparva P (2011) *J Solut Chem* 40:224–230. <https://doi.org/10.1007/s10953-010-9646-2>
43. Garkani-Nejad Z, Rashidi-Nodeh H (2010) *Electrochim Acta* 55:2597–2605. <https://doi.org/10.1016/j.electacta.2009.11.083>
44. Honarasa F, Yousefinejad S, Nasr S, Nekoeina M (2015) *J Mol Liq* 212:52–57. <https://doi.org/10.1016/j.molliq.2015.08.055>
45. Shamsipur M, Sirouejinejad A, Hemmateenejad B, Abbaspour A, Sharghi H, Alizadeh K, Arshadi S (2007) *J Electroanal Chem* 600:345–358. <https://doi.org/10.1016/j.jelechem.2006.09.006>
46. Dai YM, Liu H, Niu LL, Chen C, Chen XQ, Liu YM (2016) *J Cent South Univ* 23:1906–1914. <https://doi.org/10.1007/s11771-016-3246-2>
47. Miličević A, Novak Jovanović I (2021) *J Mol Liq* 335:116223. <https://doi.org/10.1016/j.molliq.2021.116223>
48. Touhami I, Messadi D (2019) *Energy Procedia* 157:522–532. <https://doi.org/10.1016/j.egypro.2018.11.216>
49. Kleinova M, Hewitt M, Brezova V, Madden JC, Cronin MTD, Valko M (2007) *Gen Physiol Biophys* 26:97
50. Liu H, Wen Y, Luan F, Gao Y, Li X (2009) *Cent Eur J Chem* 7:439–445. <https://doi.org/10.2478/s11532-009-0033-z>
51. Nesměrák K, Doležal R, Hudská V, Bártl J, Šticha M, Waissner K (2010) *Electroanalysis* 22:2117–2122. <https://doi.org/10.1002/elan.201000092>
52. Fatemi MH, Hadjmohammadi MR, Kamel K, Biparva P (2007) *Bull Chem Soc Jpn* 80:303–306. <https://doi.org/10.1246/bcsj.80.303>
53. Beheshti A, Riahi S, Ganjali MR (2009) *Electrochim Acta* 54:5368–5375. <https://doi.org/10.1016/j.electacta.2009.04.020>

54. Alizadeh K, Shamsipur M (2008) *J Mol Struct THEOCHEM* 862:39–43. <https://doi.org/10.1016/j.theochem.2008.04.021>
55. Namazian M, Norouzi P, Ranjbar R (2003) *J Mol Struct THEOCHEM* 625:235–241. [https://doi.org/10.1016/S0166-1280\(03\)00070-8](https://doi.org/10.1016/S0166-1280(03)00070-8)
56. Namazian M, Norouzi P (2004) *J Electroanal Chem* 573:49–53. <https://doi.org/10.1016/j.jelechem.2004.06.020>
57. Cape JL, Bowman MK, Kramer DM (2006) *Phytochemistry* 67:1781–1788. <https://doi.org/10.1016/j.phytochem.2006.06.015>
58. Gillet N, Levy B, Moliner V, Demachy I, de la Lande A (2017) *J Comput Chem* 38:1612–1621. <https://doi.org/10.1002/jcc.24802>
59. Xue ZM, Chen CH (2006) *Mol Simul* 32:401–408. <https://doi.org/10.1080/089270206006669999>
60. Riahi S, Norouzi P, Moghaddam AB, Ganjali MR, Karimipour GR, Sharghi H (2007) *Chem Phys* 337:33–38. <https://doi.org/10.1016/j.chemphys.2007.06.018>
61. Li H, Xu L, Su Q (1995) *Anal Chim Acta* 316:39–45. [https://doi.org/10.1016/0003-2670\(95\)00356-5](https://doi.org/10.1016/0003-2670(95)00356-5)
62. Nikolić S, Miličević A, Trinajstić N (2006) *Croat Chem Acta* 79:155–159
63. Bouarra N, Nadji N, Kherouf S, Nouri L, Boudjemaa A, Bachari K, Messadi D (2022) *J Turk Chem Soc A* 9:709–720. <https://doi.org/10.18596/jotcsa.1065043>
64. Weininger D (1988) *J Chem Inf Comput Sci* 28:31–36. <https://doi.org/10.1021/ci00057a005>
65. Weininger D, Weininger A, Weininger JL (1989) *J Chem Inf Comput Sci* 29:97–101. <https://doi.org/10.1021/ci00062a008>
66. Weininger D (1990) *J Chem Inf Comput Sci* 30:237–243. <https://doi.org/10.1021/ci00067a005>
67. CORAL. <http://www.insilico.eu/coral/>
68. Toropov AA, Benfenati E (2007) *Curr Drug Discov Technol* 4:77–116. <https://doi.org/10.2174/157016307781483432>
69. Toropov AA, Toropova AP, Benfenati E, Nicolotti O, Carotti A, Nesmerak K, Veselinović AM, Veselinović JB, Duchowicz PR, Bacelo D, Castro EA, Rasulev BF, Leszczynska D, Leszczynski J (2015) In: Roy K (ed) *Quantitative structure-activity relationships in drug design, predictive toxicology, and risk assessment*. IGI Global, Hershey, pp 560–585. <https://doi.org/10.4018/978-1-4666-8136-1.ch015>
70. Toropov AA, Schultz TW (2003) *J Chem Inf Comput Sci* 43:560–567. <https://doi.org/10.1021/ci025555n>
71. Toropov A, Nesmerak K, Raska I, Waisser K, Palat K (2006) *Comput Biol Chem* 30:434–437. <https://doi.org/10.1016/j.compbiolchem.2006.09.003>
72. Toropov AA, Nesmerak K (2012) *Chem Phys Lett* 539–540:204–208. <https://doi.org/10.1016/j.cplett.2012.04.061>
73. Nesmerak K, Toropov AA, Toropova AP, Kohoutova P, Waisser K (2013) *Eur J Med Chem* 67:111–114. <https://doi.org/10.1016/j.ejmech.2013.05.031>
74. Toropov A, Toropova A (2004) *J Mol Struct THEOCHEM* 711:173–183. <https://doi.org/10.1016/j.theochem.2004.10.003>
75. Toropov AA, Toropova AP (2017) *Mutat Res Genet Toxicol Environ Mutagen* 819:31–37. <https://doi.org/10.1016/j.mrgentox.2017.05.008>
76. Toropov AA, Toropova AP (2020) *Sci Total Environ* 737:139720. <https://doi.org/10.1016/j.scitotenv.2020.139720>
77. Toropov AA, Raška I, Toropova AP, Raškova M, Veselinović AM, Veselinović JB (2019) *Sci Total Environ* 659:1387–1394. <https://doi.org/10.1016/j.scitotenv.2018.12.439>
78. Fjodorova N, Vračko M, Tušar M, Jezierska A, Novič M, Kühne R, Schüürmann G (2010) *Mol Divers* 14:581–594. <https://doi.org/10.1007/s11030-009-9190-4>

Chapter 7

Building Up QSPR for Polymers Endpoints by Using SMILES-Based Optimal Descriptors



Valentin O. Kudyshkin and Alla P. Toropova

Abstract The general scheme of QSPR analysis of endpoints related to polymers is described. The basic idea of the approach is building up a model of a polymer as a mathematical function of monomer structure represented by a simplified molecular input line-entry system (SMILES). The suitability of so-called hybrid optimal descriptors in QSPR analysis of polymer systems is suggested and discussed. QSPR models for glass transition temperature and refractive index are represented in detail. Possible ways of evolution of the QSPR for polymers are listed and discussed.

Keywords Polymer · QSPR/QSAR · Monte Carlo method · SMILES · Quasi-SMILES

Abbreviations

ANN	Artificial Neural Networks
CII	Correlation Intensity Index
IIC	Index of Ideality of Correlation
MLR	Multiple Regression Analysis
PLS	Partial Least Squares regression analysis
RF	Random Forest

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-28401-4_7.

V. O. Kudyshkin
Institute of Polymer Chemistry and Physics, Academy of Sciences of the Republic of Uzbekistan,
Kodyri Street 7b, Tashkent 100128, Uzbekistan

A. P. Toropova (✉)
Laboratory of Environmental Chemistry and Toxicology, Department of Environmental Health
Science, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156
Milano, Italy
e-mail: alla.toropova@marionegri.it

QSAR	Quantitative Structure–Activity Relationships
QSPR	Quantitative Structure–Property Relationships
SMILES	Simplified Molecular Input Line-Entry System
SVM	Support Vector Machine

7.1 Introduction

It is rather difficult, perhaps, impossible, to name classes of economically useless substances. It is perhaps even more challenging to call substances that are unpromising, useless, and not of any interest from the point of view of the theory and practice of the natural sciences. According to Engels (Dialectics of Nature. Frederick Engels, 1883), “Life is the mode of existence of protein bodies”. In other words, life is the mode of existence of biopolymers. It seems that conventional polymers should also have some applications and some practical significance (Table 7.1).

The rational use of polymeric materials requires data on their physicochemical as well as biochemical properties. Experimental determination of all properties of polymers, and even more so of their solutions, alloys, and mixtures, is impossible. Under such circumstances, developing appropriate models is a promising task. If for organic, inorganic, and coordination compounds, multiple structural descriptors have been developed, based on which the corresponding models are built (boiling, melting points, solubility, toxicity), whereas for polymers, in that case, developing such descriptors is carried out according to somewhat different rules, taking into account the peculiarities of the molecular structure polymers. Often, monomer units are the basis for developing quantitative structure–property (activity) relationships (QSPR/QSAR) for different polymers. Unfortunately, this is not always possible because there are situations when different polymers consist of identical monomer units. However, owing to the significant economic and scientific sounds of polymers, many models for phenomena related to polymers cannot be reached via QSPR/QSAR analysis (Table 7.2).

It follows from the above that the models describing the behaviour of polymers are numerous and varied. Here, we consider recently proposed approaches to solving the problems of QSPR/QSAR related to polymer systems.

7.1.1 *The General Scheme of QSPR/QSAR Analysis of Endpoints Related to Polymers*

Most often, polymers’ properties are modelled using the molecular structure of monomer units [46–49]. In the case of modelling the properties of polymer solutions in organic solvents, the structures of monomer units are considered together with the molecular structure of solvents [50]. The list of physicochemical properties for which

Table 7.1 Use of polymers in economics and natural science research

Application area	Comment	References
Technology	The automotive industry	[1]
	Molecularly imprinted polymers	[2]
	Improving thermal conductivity of multi-walled carbon nanotubes	[3]
	Optics	[4]
	In the telecommunications industry, medicine, and analytical chemistry	[5]
	Glass transition temperature	[6]
	Intrinsic viscosity in polymer–solvent combinations	[7]
	Optics and mechanics properties of polymers	[8]
	Refractive index, glass transition thermal decomposition temperature, solubility	[9]
	Polymer photovoltaic research	[10]
Electronics	Superparamagnetic polyacrylamide/magnetite composite gels	[11]
	The efficiency of polymer solar cells	[10]
	Generation and transfer of energy	[10, 12]
Medicine	Drug discovery; anti-Alzheimer drugs	[3, 13]
	The aesthetic action of polymer systems	[14]
	Antimicrobial activity	[15]
	Pharmaceuticals	[16]
	Anticancer therapy	[17, 18]
Agriculture	Polymeric foams	[19–21]
	Innovative polymeric materials and intelligent delivery systems; increasing the efficiency of pesticides and herbicides; protecting the environment through filters or catalysts to reduce pollution and clean up existing pollutants	[22]
	Superabsorbent polymers	[23]
	Systems using nature-derived polymers for agriculture	[24]
	Drug delivery, bioremediation, firefighting, biosensors, food industries, thermal energy storage, and tissue engineering	[25]
Nanotechnology	The effect of nanosurfactant in emulsion polymerization	[26]
	Polymer brushes: prevention of bacterial adherence and cell protection	[27]
	Cooperative phenomena “nanoparticles-polymers”: new information and sensor technology approaches may be possible	[28]

Table 7.2 Phenomena related to polymers are objects of different manners of modelling

The object to model	Comment	References
Viscosity	High molecular weight, viscoelastic polymers are used for heavy oil recovery	[29]
	The viscoelastic properties of polymers have been widely studied due to their extensive range of engineering applications in aerospace and automotive industries, fluid transport, and electronics	[30]
	Thermo-viscoelastic shape memory polymers are an emerging class of active materials that respond to a specific temperature influenced by a shape change	[31]
Viscoelasticity	Shape memory polymers an increasing potential for various applications in biomedicine	[32–34]
Diffusion	Tune of time the release of drugs from a polymer matrix	[35]
Elastic properties	Effective properties of the composite structure for optimization of the design of composite structures	[36]
Electrical conductivity	Applications in electronics, sensors, aerospace, and shielding	[37]
Thermo-elastic properties	Thermo-plastic polymers have been widely used to fabricate engineering components in industries ranging from automotive and aerospace to biomedical fields due to their excellent impact resistance, high strength-to-weight ratio, and good bio-affinity	[38]
	Various forms of 3D printing systems rely on the extrusion of polymer materials	[39]
Biodegradation	Biodegradable polymers, mainly aliphatic polyesters, have highly desirable applications in the biomedical field and are presently being used as disposable products (e.g. syringes, blood bags), supporting materials (e.g. sutures, bone plates), artificial tissue/organs (e.g. artificial heart, kidney, eyes), and controlled release formulations for use with various drugs and hormones	[40]

(continued)

Table 7.2 (continued)

The object to model	Comment	References
Laser-based polymers	The parts are obtained without needing moulds, cutting tools, or other auxiliary resources, starting from a computer-aided design with a laser. The technology can handle details with complex geometries with excellent efficiency and near-zero material waste	[41]
Polymer-supported membranes	Polymer-supported membranes as models of the cell surface are the tools of modern genetic engineering. Bioorganic chemistry makes it possible to tune many biomolecule types to supported membranes	[42]
Smart materials	The so-called soft matters, touted to be the next generation intelligent materials, can be categorized into many different types, such as gels, shape memory polymers, dielectric elastomers, liquid crystals	[43]
Fuel cells	Fuel cells employing polymer systems are promising candidates for electric vehicle applications. The polymer electrolyte provides room temperature start-up, eliminating corrosion-related problems	[44]
Polymer solvent systems	The intrinsic viscosity of polymer solutions has technological and biomedical applications	[45]

QSPR/QSAR models are developed according to the “structure–property” paradigm includes the following: refractive indices [5, 51–53]; critical solution temperature [54, 55]; solubility [56, 57]; transport behaviour in amorphous polymeric materials [58]; solubility of CO₂ and N₂ in polymers [21]; melting point and glass transition temperature [59]; thermal decomposition [60, 61]; retention factor [62]; flammability characteristics [63]; Flory–Huggins parameter [64]; micellar properties [65]; and binding of drugs to polymer [66]. Below are some approaches to QSPR/QSAR analysis examined in more detail.

7.1.2 *QSPR Analysis of Endpoints Related to Polymers with MLR*

In QSPR studies based on multiple regression analysis (MLR), the goal is to find one or more equations that are functions of a small number of structure-based molecular descriptors that accurately predict the experimental property. As it is possible to

generate a large number of molecular descriptors for each compound in the data set, the problem becomes how to efficiently select the set of molecular descriptors that yield a reliable relationship [45].

7.1.3 QSPR/QSAR Analysis of Endpoints Related to Polymers with PLS

Quantitative structure–activity and structure–property relationship models should contain detailed information regarding how differences in the molecular structure of compounds correlate with differences in the observed biological or other physicochemical properties of those compounds. Partial least squares (PLS) regression analysis allows for identifying specific structural trends related to observed properties' differences. The study of the completed model is the last step of the process [67]. PLS models are built up with different descriptors such as Constitutional (molecular composition, molecular weight, number of atoms/bonds, number of H-bond donors/acceptors); topological (2D structural formula, Kier–Hall indices, branching); geometrical (3D structure of molecule, molecular volume, polar and non-polar surface area); electrostatic (charge distribution, atomic partial charges, electronegativity); and quantum mechanical (electronic structure, HOMO–LUMO energies, dipole moment) [68].

7.1.4 QSPR Analysis of Endpoints Related to Polymers with ANN

QSPR models can be constructed to predict polymer properties using artificial neural networks (ANN). ANN-Procedures are carried out using functional monomers, which serve as input for generating molecular descriptors. Constitutional, topological, geometrical, electrostatic, and quantum mechanical descriptors are suitable for building ANN models. Some sets of descriptors fed to ANN should be selected preliminary as input vectors. As a rule, the networks consist of an input layer, an output layer, and some number of intermediate layers known as hidden layers. Each unit in the network is influenced by those units to which it is connected. The degree of influence is dictated by the values of the links or connections. The system's overall behaviour can be modified by adjusting the importance of the relationships or weights through a repeated application of a learning algorithm. The advantage of the ANN approach is the possibility of building models for nonlinear phenomena [69].

7.1.5 QSPR Analysis of Endpoints Related to Polymers with SVM

A support vector machine (SVM) is an original and capable classification and regression method. SVM models are primarily developed for classification problems; however, they can also be applied to solve nonlinear regression problems. To realize an accurate regression model, SVM is used to construct a nonlinear model based on a subset of descriptors. The performances of SVM for regression rely on the combination of several parameters [70].

7.2 Significant Notes

The four approaches discussed regarding modelling the properties of polymers do not exhaust all the ideas related to this topic. However, they are currently the most common.

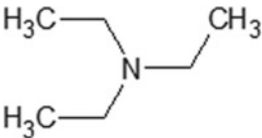
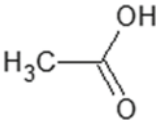
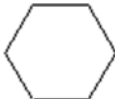
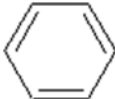
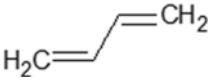
Speaking of QSPR/QSAR, it is necessary to take into account that all proposed models must comply with the five famous OECD principles, which state:

- A defined endpoint;
- An unambiguous algorithm;
- A defined applicability domain;
- Appropriate measures of goodness-of-fit and robustness;
- A mechanistic interpretation, if possible.

7.3 Building Up Models of Polymers Endpoints Using SMILES

The approaches considered above are widely used to construct models of polymer systems' physicochemical and biochemical behaviour and throughout the whole area of QSPR/QSAR analysis. However, all of the approaches mentioned need to use various additional descriptors. This section discusses methods that require only data on the structure of monomeric units (without additional geometrical, electrostatic, and quantum mechanical descriptors) for their implementation. The generalized name for these approaches is formulated as "optimal descriptors" calculated via a simplified molecular input line-entry system (SMILES) [71].

Table 7.3 Representation of molecular structure by SMILES

Name	Structure	SMILES
Hydrogen cyanide	$\text{HC}\equiv\text{N}$	<chem>C#N</chem>
Triethylamine		<chem>CCN(CC)CC</chem>
Acetic acid		<chem>CC(=O)O</chem>
Cyclohexane		<chem>C1CCCCC1</chem>
Benzene		<chem>c1ccccc1</chem>
1,3-Butadiene		<chem>C=CC=C</chem>

7.3.1 SMILES

For the practical implementation of input data in the form of SMILES, as practice shows, it is better to use programs that use not only capital letters of the Latin alphabet but also small ones to point out aromaticity. The ACD/ChemSketch (www.acdlabs.com) is an example of such a program. SMILES can be used to represent chemical reactions, but these features, while in demand for database development, have not yet been used in QSPR/QSAR analysis. Table 7.3 contains examples of the representation of substances via SMILES.

7.3.2 Optimal SMILES-Based Descriptors

The set of descriptors was calculated using the adjacency matrix, which was the source of the vertex degrees. The vertex degree for the k th vertex is actually the sum of the elements of the adjacency matrix in the k th row (or column). It has been shown

that the correlation potential of a descriptor can be improved by using the optimal non-zero values of the diagonal elements in the adjacency matrix for heteroatoms (non-carbon atoms), calculating the degrees of vertices as the sum of the elements of the adjacency matrix [72].

Instead of the adjacency matrix, you can use SMILES, choosing a special correlation weight for each SMILES atom. A SMILES atom is a single character in the string SMILES (e.g. "C", "O", "N", etc.) or several characters that cannot be considered separately (e.g. "Cl", "Br", etc.). Hence, the optimal SMILES-based descriptor calculated as

$$\text{DCW}(T, N) = \sum \text{CW}(S_k) \quad (7.1)$$

The $\text{CW}(S_k)$ is the correlation weight of a SMILES atom. The Monte Carlo optimization procedure calculates the $\text{CW}(S_k)$ numerical data.

7.3.3 The Monte Carlo Optimization Procedure

It is assumed that as a model, there is a one-parameter equation (QSPR-regression model) of the form [73]:

$$\text{Endpoint} = C_0 + C_1 \times \text{DCW}(T, N) \quad (7.2)$$

C_0 and C_1 are the regression coefficients; T is the threshold, i.e. an integer limits frequency of a SMILES atom in the training set; and N is the number of epochs of the optimization process (step-by-step modifications) of all correlation weights accepted to build up the model accordingly to the threshold (T).

7.3.4 The Classic Scheme of Building Up the QSPR/QSAR Model Using the Optimal Descriptors

The essence of the classical model building scheme is to establish a correlation between the optimal descriptor and a property for the training set in the hope that this correlation will be preserved for similar external molecules not taken into account when building this correlation. Thus, it should be emphasized: that the classical optimization scheme is reduced to selecting such correlation weights that give the maximum value of the coefficient of determination between the endpoint and the descriptor for the entire training set. However, it was stated that the so-called balance of correlations gives more reliable models than the classical scheme [73].

7.3.5 *The Balance of Correlations for the QSPR/QSAR Model Using the Optimal Descriptors*

The balance of correlations [73] provides the division of data for building models into two groups—active and passive training samples. The molecules of the active training set are used to build the model. Passive training set molecules are used to control whether the resulting model works for “outside observers” (passive training set). In practice, this is achieved by modifying the objective function (instead of the determination coefficient on the whole training set) for optimization to the following:

$$TF_0 = r_{AT} + r_{PT} - |r_{AT} - r_{PT}| \times 0.1 \quad (7.3)$$

The r_{AT} and r_{PT} are correlation coefficients between the observed and predicted endpoints for the active and passive training sets, respectively.

7.3.6 *Search and Use for Reliable Criteria of the Predictive Potential of QSPR/QSAR Models Based on the Optimal Descriptors*

The index of ideality of correlation (IIC) [74] and the correlation intensity index (CII) [75] are two relatively new criteria for the predictive potential of the QSPR/QSAR models. The IIC_C is calculated with data on the calibration set as the following:

$$IIC_C = r_C \frac{\min(-MAE_C, +MAE_C)}{\max(-MAE_C, +MAE_C)} \quad (7.4)$$

$$\min(x, y) = \begin{cases} x, & \text{if } x < y \\ y, & \text{otherwise} \end{cases} \quad (7.5)$$

$$\max(x, y) = \begin{cases} x, & \text{if } x > y \\ y, & \text{otherwise} \end{cases} \quad (7.6)$$

$$-MAE_C = \frac{1}{-N} \sum |\Delta_k|, \quad -N \text{ is the number of } \Delta_k < 0 \quad (7.7)$$

$$+MAE_C = \frac{1}{+N} \sum |\Delta_k|, \quad +N \text{ is the number of } \Delta_k \geq 0 \quad (7.8)$$

$$\Delta_k = \text{observed}_k - \text{calculated}_k \quad (7.9)$$

The observed and calculated are corresponding values of the endpoint.

The correlation intensity index (CII), similarly to the above IIC, was developed as a tool to improve the quality of the Monte Carlo optimization to build up QSPR models.

The CII_C calculated as follows:

$$CII_C = 1 - \sum \text{Protest}_k \quad (7.10)$$

$$\text{Protest}_k = \begin{cases} R_k^2 - R^2, & \text{if } R_k^2 - R^2 > 0 \\ 0, & \text{otherwise} \end{cases} \quad (7.11)$$

The R^2 is the determination coefficient for a set that contains n substances. The R_k^2 is the determination coefficient for $n - 1$ substances of a group after removing of k th substance. Hence, if the $(R_k^2 - R^2)$ is more significant than zero, the k th substance is an “opponentist” for the correlation between experimental and predicted values of the set. A small sum of “protests” means a more “intensive” correlation.

7.3.7 Hybrid Optimal Descriptors

Both a molecular graph and a SMILES are a representation of a molecular structure. These representations partly coincide (that is, they contain identical information) and somewhat differ (complement each other). Optimal descriptors calculated from the correlation weights of molecular features extracted from SMILES and molecular features extracted from graphs are called hybrid optimal descriptors [76]. Using hybrid optimal descriptors can improve the statistical characteristics of a model.

7.3.8 Model Complication

Another way that can lead to model improvement is the complication of optimal descriptors [51]. The SMILES components of the optimal descriptor can connect to the calculation scheme, considering the influence of neighbouring pairs of SMILES atoms and neighbouring triplets of SMILES atoms. In addition, it is possible to involve global SMILES attributes, such as the configuration of covalent bonds, as well as configurations of four atoms (nitrogen, oxygen, sulphur, phosphorus and/or fluorine, chlorine, bromine, iodine). For graph components of optimal descriptors, complication can be achieved through correlation weighting the sums and differences of various graph invariants. However, it should be noted that increasing the complexity of the model often leads to a significant improvement in the statistical quality of the model for the training set, but which is accompanied by deterioration of the statistical quality of the model for the test set.

7.4 Examples of Improving Models Built Up with Optimal Descriptors

The described possibilities for improving models calculated using optimal descriptors should be confirmed with specific examples. Below are the results of using the correlation intensity index (CII) to increase the efficiency of the Monte Carlo method for developing models for glass transition temperature and refractive index.

7.4.1 Development of a New Conception to Building Up a Model

The application of the CII is the main improvement of the models for glass transition temperature and refractive index of polymers. The optimal descriptor has been defined as the following:

$$\text{DCW}(T^*, N^*) = \sum \text{CW}(S_k) + \sum \text{CW}(SS_k) + \sum \text{CW}(SSS_k) \quad (7.12)$$

S_k is the SMILES atom (one symbol or a group of symbols which cannot be examined separately); SS_k and SSS_k are two and three connected SMILES atoms. $\text{CW}(S_k)$, $\text{CW}(SS_k)$, and $\text{CW}(SSS_k)$ are the correlation weights for the attributes mentioned above of the SMILES. The correlation weights were calculated by the Monte Carlo optimization with the target function calculated as the following:

$$\text{TF} = \text{TF}_0 + 0.5 \times \text{IIC} + 0.5 \times \text{CII} \quad (7.13)$$

Models for the glass transition temperature GTT (experimental data taken [77]):

$$\text{GTT}'\text{K} = 303.15 (\pm 2.50) + 6.536 (\pm 0.225) * \text{DCW}(1, 15) \quad (7.14)$$

Models for the refractive index RI (experimental data taken [78]):

$$\text{RI} = 1.5009 (\pm 0.0007) + 0.00427 (\pm 0.00007) * \text{DCW}(1, 15) \quad (7.15)$$

It is to be noted both models were obtained by the same calculating scheme. Figure 7.1 contains the graphical interface of the CORAL method (<http://www.insilico.eu/coral>) for the calculations above endpoints.

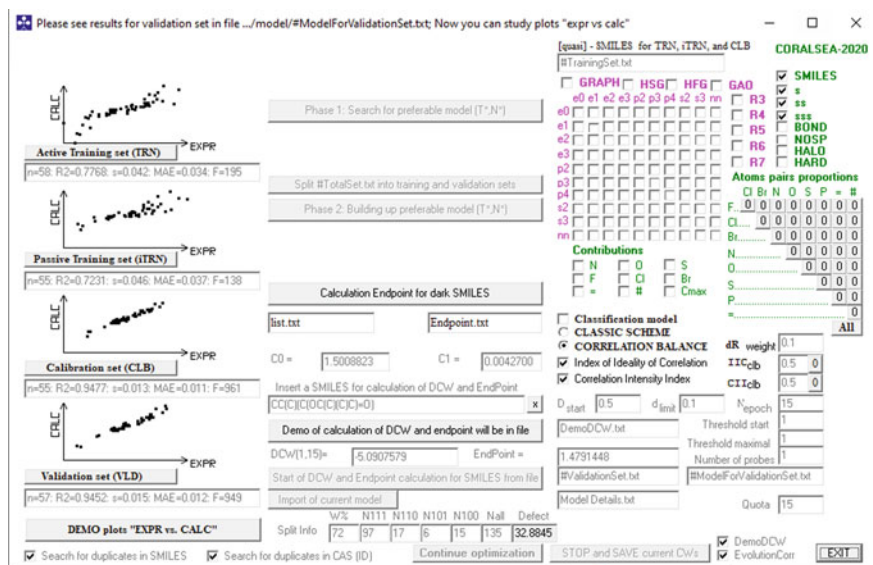


Fig. 7.1 Graphical representation of the CORAL method applied to build up models for glass transition temperature and refractive index

7.4.2 QSPR Models for the Glass Transition Temperature

The best model for the glass transition temperature built up with the optimal SMILES-based descriptors [77] is characterized by a determination coefficient of 0.9058. The model was calculated with hybrid optimal descriptors sensitive combinations of single SMILES atoms together with their connected pairs and three SMILES atoms. In addition, the model is sensitive to the presence of Morgan's extended connectivity of first- and second-order as well as to the fact of five- and six-member rings [77]. The Monte Carlo optimization was carried out by considering the IIC values. Monte Carlo optimization of the descriptor based on the mentioned SMILES combinations of atoms, without taking into account molecular graph invariants, in the case of using the CII, gives a model for the glass transition temperature for the case of the same polymers, which is characterized (validation set) by a coefficient of determination of 0.9184. Thus, thanks to using CII in Monte Carlo calculations, obtaining a glass transition temperature model with improved predictive potential is obtained. It is a more straightforward model calculated using only SMILES without involving molecular graph invariants. *Supplementary materials* section contains technical details on the model (Table S1 for QSPR models for the glass transition temperature).

7.4.3 QSPR Models for the Refractive Index

The best model for the refractive index that has been built up with the optimal SMILES-based descriptors [78] is characterized by a determination coefficient of 0.9028. This model is calculated by means of a correlation adjustment of a descriptor that includes both molecular features expressed by SMILES attributes and molecular features represented by invariants of molecular graphs. Models suggested in the literature [78] were built up by the Monte Carlo technique by applying the IIC. The approach described above (the same descriptor and the same Monte Carlo optimization that involves CII) gives for the refractive index model, which is statistically characterized (validation set) by the determination coefficient of 0.9452. Thus, a simpler model with improved predictive potential was also obtained for modelling the refractive index. *Supplementary materials* section contains technical details on the model (Table S2 for QSPR models for refractive index).

7.5 Comparison QSPR-Models

The comparison of the statistical quality of different models and models suggested here confirms that models obtained with the Monte Carlo optimization involving IIC and CII characterize models with quite comparable statistical quality (Table 7.4).

Table 7.4 Comparison of the statistical quality for different models

Endpoint	Method	Statistical quality	References
Glass transition temperature	CODESSA	$R^2 = 0.946$	[46]
	CORAL	Training set $R^2 = 0.7477$, validation set $R^2 = 0.9058$	[77]
	MLR	$R^2 = 0.755$	[79]
	SVM	$R^2 = 0.479$	[79]
	RF	$R^2 = 0.721$	[79]
	ANN	Training set $R^2 = 0.8477$, test set $R^2 = 0.5272$	[80]
	DRAGON	PLS: $R^2 = 0.848$ SVM: $R^2 = 0.886$ MLR: $R^2 = 0.860$ Least absolute shrinkage and selection operator: $R^2 = 0.869$ Elastic net: $R^2 = 0.880$ Gaussian process regression: $R^2 = 0.899$	[81]

(continued)

Table 7.4 (continued)

Endpoint	Method	Statistical quality	References
	Virtual Computational Chemistry Laboratory	Training set $R^2 = 0.9473$, validation set $R^2 = 0.9283$	[82]
	SVM	Training set $R^2 = 0.920$, validation set $R^2 = 0.779$	[83]
	CORAL	Training set $R^2 = 0.683$, validation set $R^2 = 0.877$	[84]
	CORAL	Training set $R^2 = 0.4490$, validation set $R^2 = 0.9184$	In this work
Refractive index	CODESSA	$R^2 = 0.940$	[47]
	MLRA	$R^2 = 0.929$	[49]
	Correlating the refractive indices with two 2D descriptors	$R^2 = 0.801$	[49]
	Correlation between the refractive indices and the three 2D descriptors	$R^2 = 0.918$	[49]
	Regression model based on DRAGON and CORAL descriptors	Training set $R^2 = 0.96$, validation set $R^2 = 0.95$	[51]
	ANN	Training set $R^2 = 0.971$, validation set $R^2 = 0.9613$	[52]
	PLS with DRAGON and PaDEL descriptors	Training set $R^2 = 0.895$, validation set $R^2 = 0.707$ Training set $R^2 = 0.899$, validation set $R^2 = 0.794$ Training set $R^2 = 0.897$, validation set $R^2 = 0.766$ Training set $R^2 = 0.896$, validation set $R^2 = 0.796$	[53]
	CORAL	Training set $R^2 = 0.7764$, validation set $R^2 = 0.9028$	[78]
	DRAGON	Training set $R^2 = 0.907$, validation set $R^2 = 0.823$	[85]
	Genetic algorithm and QSARINS	Training set $R^2 = 0.932$, validation set $R^2 = 0.882$	[86]
	CORAL	Training set $R^2 = 0.7788$, validation set $R^2 = 0.9452$	In this work

It should be noted that in the case of using the objective function of the involving CII, as well as in the case of the objective function of the involving Monte Carlo IIC, optimization improves the statistical quality of the models for the calibration set as well for the validation set, but to the detriment of the training set.

7.6 Possible Ways of Evolution of the QSPR for Polymers

The list of main unpleasant peculiarities of QSPR/QSAR is as follows: (i) possibility of “chance correlations”; (ii) possibility of overtraining; (iii) possibility of weak reproducibility of statistical quality of an approach suggested [87].

Often the modern QSPR/QSAR researches are based solely on one distribution of available data into the training and validation sets. According to many authors, a rational split into training and validation sets gives better statistical results for the validation sets than models based on a group of random splits. An examination of several splits decreases the probability of “chance correlations”: solely one good correlation easily can become a chance correlation; however, three good correlations hardly can be “chance correlations”.

The number of statistical characteristics aimed to measure the predictive potential of a model gradually increases, despite the attractiveness of a small number of criteria for the predictive potential for practical applications. On the one hand, the diversity of different standards of predicting potential is a tool to improve the quality of QSPR/QSAR models. On the other hand, this situation causes uncertainty in choosing the best model. In other words, contradictions in the recommendations of various criteria force the researcher to search for the best choice in a maze of numerous possibilities.

As a rule, the contribution of the molecular structure is crucial to an endpoint. However, any physicochemical property, as well as any biological activity, is a mathematical function of many different conditions and circumstances. In other words, non-equilibrium physicochemical processes or pharmaceutical effects are caused by not only molecular structure but also physicochemical conditions (e.g. temperature, humidity) and circumstances (noise/silence, illumination/darkness). Apparently, one can agree with the above postulation, but the majority of QSPR/QSAR has built up without taking into account something besides molecular structure. However, it is to be noted that in some cases, the molecular structure is not informative to build up a predictive model of endpoints, e.g. endpoints related to polymers and/or nano-materials. Sometimes, in addition to the molecular structure, one should consider experimental conditions. Thus, the definition of a model as a mathematical function of experimental conditions (after consultations with experimentalists) could be a shorter and consequently more attractive way to solve the corresponding tasks.

QSPR/QSAR should be assessed as a surrogate of a real experiment for traditional substances as well for polymers. QSPR/QSAR aimed to measure an endpoint value. However, to expect adequate prediction of physicochemical and biochemical behaviour of an arbitrary substance by means of the QSPR/QSAR model is naive. Despite the above-mentioned thesis, QSPR/QSAR has become an integral part of modern science as a tool to detect “fuzzy tendencies” in the behaviour of groups of substances. This fact logically echoes the theory of fuzzy sets. This is not surprising, as fuzzy set theory has solved some QSPR/QSAR analysis problems. One can extract two components in the total wide variety of QSPR/QSAR studies: (i) “extensive” studies and (ii) “intensive” studies. “Extensive” studies aim to integrate the results of

applying current approaches to solve practical tasks. The “intensive” studies attempt to develop new conceptions of the QSPR/QSAR analysis. Naturally, a small part of the results of the “intensive” studies gradually become a tool for robust “extensive” studies. Nowadays, multi-target QSPR/QSAR is a part of “intensive” studies. The development of criteria for models’ predictive potential is also a part of the “intensive” studies. Maybe searching for the similarity of endpoints will also become part of “intensive” QSPR/QSAR research.

Reliable prediction of endpoints related to different substances using unambiguous algorithms is an attractive alternative to experimental investigation.

7.7 Quasi-SMILES Can Be a Tool for the Discussion of Experimentalists and Model Developers

Quasi-SMILES is a sequence of symbols representing all available eclectic data, i.e. the molecular structure and different conditions, which can influence examined endpoint [88, 89]. Descriptor calculated with optimal correlation weights of different fragments of quasi-SMILES defined by the Monte Carlo technique is used to predict an endpoint as a mathematical function of molecular structure and arbitrary experimental conditions. The statistical quality of the models based on correlation weights of fragments of quasi-SMILES can be better than the statistical quality of models obtained with traditional SMILES.

7.8 Conclusions

The QSPR for polymers can be developed from SMILES representing the molecular structure of monomer units or their compositions. These models can be improved by means of applying the IIC and CII. Quasi-SMILES is a possible way to establish new models which will extract all available eclectic data on the endpoint of interest. All QSPR/QSAR related to polymer systems should be qualified as random events.

References

1. Creton B, Veyrat B, Klopffer M-H (2022) Fluid Phase Equilib 556:113403. <https://doi.org/10.1016/j.fluid.2022.113403>
2. Lowdon JW, Ishikura H, Kvernenes MK, Caldara M, Cleij TJ, van Grinsven B, Eersels K, Diliën H (2021) Computation 9(10):103. <https://doi.org/10.3390/computation9100103>
3. Qu Z, Wang K, Xu C-A, Li Y, Jiao E, Chen B, Meng H, Cui X, Shi J, Wu K (2021) Chem Eng J 421:129729. <https://doi.org/10.1016/j.cej.2021.129729>
4. Owolabi TO, Abd Rahman MA (2021) Polymers 13(16):2697. <https://doi.org/10.3390/polym13162697>

5. Schustik SA, Cravero F, Ponzoni I, Díaz MF (2021) *Comput Mater Sci* 194:110460. <https://doi.org/10.1016/j.commatsci.2021.110460>
6. Miccio LA, Schwartz GA (2021) *Macromolecules* 54(4):1811–1817. <https://doi.org/10.1021/acs.macromol.0c02594>
7. Wang S, Cheng M, Zhou L, Dai Y, Dang Y, Ji X (2021) *SAR QSAR Environ Res* 32(5):379–393. <https://doi.org/10.1080/1062936X.2021.1902387>
8. Minami T, Okuno Y (2018) *MRS Adv* 3(49):2975–2980. <https://doi.org/10.1557/adv.2018.454>
9. Venkatraman V, Alsberg BK (2018) *Polymers* 10(1):103. <https://doi.org/10.3390/POLYMI0010103>
10. Bernardo G, Deb N, King SM, Bucknall DG (2016) *J Polym Sci Part B Polym Phys* 54(10):994–1001. <https://doi.org/10.1002/polb.24002>
11. Filho E, Brito E, Silva R, Streck L, Bohn F, Fonseca J (2021) *J Dispers Sci Technol* 42(10):1504–1512. <https://doi.org/10.1080/01932691.2020.1774382>
12. Yu T, Kim R, Park H, Yi J, Kim W-S (2014) *Chem Phys Lett* 592:265–271. <https://doi.org/10.1016/j.cplett.2013.12.006>
13. Ghasemi G (2019) *J Sci Ind Res* 78(5):323–327
14. Golubović M, Lazarević M, Zlatanović D, Krtinić D, Stoičkov V, Mladenović B, Milić DJ, Sokolović D, Veselinović AM (2018) *Comput Biol Chem* 75:32–38. <https://doi.org/10.1016/j.compbiolchem.2018.04.009>
15. Faya M, Kalthapure RS, Kumalo HM, Waddad AY, Omolo C, Govender T (2018) *J Drug Deliv Sci Technol* 44:153–171. <https://doi.org/10.1016/j.jddst.2017.12.010>
16. Podrazka M, Bącznyńska E, Kundys M, Jeleń PS, Nery EW (2017) *Biosensors* 8(1):3. <https://doi.org/10.3390/bios8010003>
17. Yipel M, Ghica MV, Albu Kaya MG, Spoiala A, Radulescu M, Ficai D, Ficai A, Bleotu C, Cornelia N (2016) *Curr Org Chem* 20(28):2934–2948. <https://doi.org/10.2174/1385272820666160919112919>
18. Yan R, Hallam A, Stockley PG, Boyes J (2014) *Biochem J* 461(1):1–13. <https://doi.org/10.1042/BJ20140173>
19. Toropov AA, Toropova AP, Begum S, Achary PGR (2016) *SAR QSAR Environ Res* 27(4):293–301. <https://doi.org/10.1080/1062936X.2016.1172666>
20. Valenzuela LM, Knight DD, Kohn J (2016) *Int J Biomater* 2016:6273414. <https://doi.org/10.1155/2016/6273414>
21. Golzar K, Amjad-Iranagh S, Modarress H (2013) *Measurement* 46(10):4206–4225. <https://doi.org/10.1016/j.measurement.2013.08.012>
22. Puoci F, Iemma F, Spizzirri UG, Cirillo G, Curcio M, Picci N (2008) *Am J Agric Biol Sci* 3(1):299–314. <https://doi.org/10.3844/ajabssp.2008.299.314>
23. Chen P, Zhang W, Luo W, Fang Y (2004) *J Appl Polym Sci* 93(4):1748–1755. <https://doi.org/10.1002/app.20612>
24. Sampathkumar K, Tan KX, Loo SCJ (2020) *iScience* 23(5):101055. <https://doi.org/10.1016/j.isci.2020.101055>
25. Behera S, Mahanwar PA (2020) *Polym-Plast Technol Mater* 59(4):341–356. <https://doi.org/10.1080/25740881.2019.1647239>
26. Paul DR, Robeson LM (2008) *Polymer* 49(15):3187–3204. <https://doi.org/10.1016/j.polymer.2008.04.017>
27. Ayres N (2010) *Polym Chem* 1(6):769–777. <https://doi.org/10.1039/b9py00246d>
28. Fujiki M, Koe JR, Terao K, Sato T, Teramoto A, Watanabe J (2003) *Polym J* 35(4):297–344. <https://doi.org/10.1295/polymj.35.297>
29. Azad MS, Trivedi JJ (2019) *Fuel* 235:218–226. <https://doi.org/10.1016/j.fuel.2018.06.030>
30. Meng R, Yin D, Drapaca CS (2019) *Int J Non-Linear Mech* 113:171–177. <https://doi.org/10.1016/j.ijnonlinmec.2019.04.002>
31. Nguyen TD, Jerry Qi H, Castro F, Long KN (2008) *J Mech Phys Solids* 56(9):2792–2814. <https://doi.org/10.1016/j.jmps.2008.04.007>
32. Li Y, Liu Z (2018) *Polymer* 143:298–308. <https://doi.org/10.1016/j.polymer.2018.04.026>

33. Omidian H, Hashemi SA, Sammes PG, Meldrum I (1998) *Polymer* 39(26):6697–6704. [https://doi.org/10.1016/S0032-3861\(98\)00095-0](https://doi.org/10.1016/S0032-3861(98)00095-0)
34. Richbourg NR, Peppas NA (2020) *Prog Polym Sci* 105:101243. <https://doi.org/10.1016/j.propolymsci.2020.101243>
35. Masaro L, Zhu XX (1999) *Prog Polym Sci* 24(5):731–775. [https://doi.org/10.1016/S0079-6700\(99\)00016-7](https://doi.org/10.1016/S0079-6700(99)00016-7)
36. Raju B, Hiremath SR, Roy Mahapatra D (2018) *Compos Struct* 204:607–619. <https://doi.org/10.1016/j.compstruct.2018.07.125>
37. Zare Y, Rhee KY (2018) *Compos Sci Technol* 155:252–260. <https://doi.org/10.1016/j.compscitech.2017.10.007>
38. Shen F, Kang G, Lam YC, Liu Y, Zhou K (2019) *Int J Plast* 121:227–243. <https://doi.org/10.1016/j.jplas.2019.06.003>
39. Duty C, Ajinjeru C, Kishore V, Compton B, Hmeidat N, Chen X, Liu P, Hassen AA, Lindahl J, Kunc V (2018) *J Manuf Process* 35:526–537. <https://doi.org/10.1016/j.jmapro.2018.08.008>
40. Sevim K, Pan J (2018) *Acta Biomater* 66:192–199. <https://doi.org/10.1016/j.actbio.2017.11.023>
41. Brighenti R, Cosma MP, Marsavina L, Spagnoli A, Terzano M (2021) *J Mater Sci* 56(2):961–998. <https://doi.org/10.1007/s10853-020-05254-6>
42. Tanaka M, Sackmann E (2005) *Nature* 437(7059):656–663. <https://doi.org/10.1038/nature04164>
43. Huang R, Zheng S, Liu Z, Ng TY (2020) *Int J Appl Mech* 12(2):2050014. <https://doi.org/10.1142/S1758825120500143>
44. Springer TE, Zowodzinski TA, Gottesfeld S (1991) *J Electrochem Soc* 138(8):2334–2342. <https://doi.org/10.1149/1.2085971>
45. Gharagheizi F (2007) *Comput Mater Sci* 40(1):159–167. <https://doi.org/10.1016/j.commatsci.2006.11.010>
46. Katritzky AR, Sild S, Lobanov V, Karelson M (1998) *J Chem Inf Comput Sci* 38(2):300–304. <https://doi.org/10.1021/ci9700687>
47. Katritzky AR, Sild S, Karelson M (1998) *J Chem Inf Comput Sci* 38(6):1171–1176. <https://doi.org/10.1021/ci980087w>
48. Cypcar CC, Camelio P, Lazzeri V, Mathias LJ, Waegell B (1996) *Macromolecules* 29(27):8954–8959. <https://doi.org/10.1021/ma961170s>
49. Xu J, Chen B, Zhang Q, Guo B (2004) *Polymer* 45(26):8651–8659. <https://doi.org/10.1016/j.polymer.2004.10.057>
50. Afantitis A, Melagraki G, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2006) *Polymer* 47(9):3240–3248. <https://doi.org/10.1016/j.polymer.2006.02.060>
51. Duchowicz PR, Fioressi SE, Baceo DE, Saavedra LM, Toropova AP, Toropov AA (2015) *Chemom Intell Lab Syst* 140:86–91. <https://doi.org/10.1016/j.chemolab.2014.11.008>
52. Xu J, Liang H, Chen B, Xu W, Shen X, Liu H (2008) *Chemom Intell Lab Syst* 92(2):152–156. <https://doi.org/10.1016/j.chemolab.2008.02.006>
53. Khan PM, Rasulev B, Roy K (2018) *ACS Omega* 3(10):13374–13386. <https://doi.org/10.1021/acsomega.8b01834>
54. Xu J, Liu L, Xu W, Zhao S, Zuo D (2007) *J Mol Graph Model* 26(1):352–359. <https://doi.org/10.1016/j.jmglm.2007.01.004>
55. Melagraki G, Afantitis A, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2007) *J Mol Model* 13(1):55–64. <https://doi.org/10.1007/s00894-006-0125-z>
56. Yu X, Wang X, Wang H, Li X, Gao J (2006) *QSAR Comb Sci* 25(2):156–161. <https://doi.org/10.1002/qsar.200530138>
57. Koç DT, Koç ML (2015) *Chemom Intell Lab Syst* 144:122–127. <https://doi.org/10.1016/j.chemolab.2015.04.005>
58. Tokarski JS, Hopfinger AJ, Hobbs JD, Ford DM, Faulon J-LM (1997) *Comput Theor Polym S7(3–4):199–214. https://doi.org/10.1016/S1089-3156(98)00007-5*
59. Bertinetto C, Duce C, Micheli A, Solaro R, Starita A, Tiné MR (2009) *J Mol Graph Model* 27(7):797–802. <https://doi.org/10.1016/j.jmglm.2008.12.001>

60. Ajloo D, Sharifian A, Behniafar H (2008) Bull Korean Chem Soc 29(10):2009–2016. <https://doi.org/10.5012/bkcs.2008.29.10.2009>
61. Mallakpour S, Hatami M, Golmohammadi H (2013) Polym Bull 70(2):715–732. <https://doi.org/10.1007/s00289-013-0906-3>
62. Porobić I, Kontrec D, Šoškić M (2013) Bioorg Med Chem 21(3):653–659. <https://doi.org/10.1016/j.bmc.2012.11.048>
63. Parandekar PV, Browning AR, Prakash O (2015) Polym Eng Sci 55(7):1553–1559. <https://doi.org/10.1002/pen.24093>
64. Xu J, Liu H, Li W, Zou H, Xu W (2008) Macromol Theory Simul 17(9):470–477. <https://doi.org/10.1002/mats.200800063>
65. Wu W, Zhang R, Peng S, Li X, Zhang L (2016) Chemom Intell Lab Syst 157:7–15. <https://doi.org/10.1016/j.chemolab.2016.06.011>
66. Brew CT, Blake JF, Mistry A, Liu F, Carreno D, Madsen D, Mu YQ, Mayo M, Stahl W, Matthews D, Maclean D, Harrison S (2018) Pharm Res 35(4):89. <https://doi.org/10.1007/s11095-018-2356-y>
67. Stanton DT (2012) Curr Comput Aided Drug Des 8(2):107–127. <https://doi.org/10.2174/157340912800492357>
68. Kholodovych V, Smith JR, Knight D, Abramson S, Kohn J, Welsh WJ (2004) Polymer 45(22):7367–7379. <https://doi.org/10.1016/j.polymer.2004.09.002>
69. Yu X, Yi B, Liu F, Wang X (2008) React Funct Polym 68(11):1557–1562. <https://doi.org/10.1016/j.reactfunctpolym.2008.08.009>
70. Mallakpour S, Hatami M, Khooshechin S, Golmohammadi H (2014) J Therm Anal Calorim 116(2):989–1000. <https://doi.org/10.1007/s10973-013-3587-0>
71. Weininger D (1988) J Chem Inf Comput Sci 28(1):31–36. <https://doi.org/10.1021/ci00057a005>
72. Randić M (1991) J Comput Chem 12(8):970–980. <https://doi.org/10.1002/jcc.540120810>
73. Toropov AA, Toropova AP, Benfenati E (2009) Int J Mol Sci 10(7):3106–3127. <https://doi.org/10.3390/ijms10073106>
74. Toropova AP, Toropov AA (2019) Mol Inform 38(8–9):1800157. <https://doi.org/10.1002/minf.201800157>
75. Toropova AP, Toropov AA (2020) Fuller Nanotub Carbon Nanostruct 28(11):900–906. <https://doi.org/10.1080/1536383X.2020.1779705>
76. Toropova AP, Toropov AA (2017) Toxicol Lett 275:57–66. <https://doi.org/10.1016/j.toxlet.2017.03.023>
77. Toropov AA, Toropova AP, Kudyshkin VO, Bozorov NI, Rashidova SS (2020) Struct Chem 31(5):1739–1743. <https://doi.org/10.1007/s11224-020-01588-8>
78. Toropov AA, Toropova AP, Kudyshkin VO (2022) Struct Chem 33(2):617–624. <https://doi.org/10.1007/s11224-021-01875-y>
79. Lee FL, Park J, Goyal S, Qaroush Y, Wang S, Yoon H, Rammohan A, Shim Y (2021) Polymers 13(21):3653. <https://doi.org/10.3390/polym13213653>
80. Epure E-L, Oniciuc SD, Hurduc N, Drăgoi EN (2021) Polymers 13(23):4151. <https://doi.org/10.3390/polym13234151>
81. Taniwaki H, Kaneko H (2022) Polym Eng Sci. First published: 24 June 2022. <https://doi.org/10.1002/pen.26058>
82. Mercader AG, Duchowicz PR (2016) Mater Chem Phys 172:158–164. <https://doi.org/10.1016/j.matchemphys.2016.01.057>
83. Higuchi C, Horvath D, Marcou G, Yoshizawa K, Varnek A (2019) ACS Appl Polym Mater 1(6):1430–1442. <https://doi.org/10.1021/ACSAPM.9B00198>
84. Toropova AP, Toropov AA, Leszczynska D, Leszczynski J (2018) J Polym Res 25:221–227. <https://doi.org/10.1007/s10965-018-1618-z>
85. Erickson ME, Ngongang M, Rasulev B (2020) Molecules 25(17):25173772. <https://doi.org/10.3390/molecules25173772>
86. Jabeen F, Chen M, Rasulev B, Ossowski M, Boudjouk P (2017) Comput Mater Sci 137:215–224. <https://doi.org/10.1016/j.commatsci.2017.05.022>

87. Toropov AA, Toropova AP (2020) *Molecules* 25(6):25061292. <https://doi.org/10.3390/molecules25061292>
88. Toropov AA, Rallo R, Toropova AP (2015) *Curr Top Med Chem* 15(18):1837–1844. <https://doi.org/10.2174/1568026615666150506152000>
89. Achary PGR, Begum S, Toropova AP, Toropov AA (2016) *Mater Discov* 5:22–28. <https://doi.org/10.1016/j.md.2016.12.003>

Part IV
Quasi-SMILES for QSPR/QSAR

Chapter 8

Quasi-SMILES-Based QSPR/QSAR Modeling



Shahin Ahmadi and Neda Azimi

Abstract Quantitative structure–property/activity relationships (QSPRs/QSARs) have been used to predict the physicochemical property and biological activity of different substances, considering that the physicochemical property/biological activity of a new or untested substance can be inferred from the molecular structure or other properties of similar compounds whose properties/activities have already been assessed. Traditional QSPR/QSAR models based on physicochemical properties and molecular information are not so successful in predicting endpoint of substances such as nanomaterials due to scarcity of available dataset in same conditions. A new approach using eclectic information as descriptors to predict the endpoint of substance materials was developed in CORAL software (<http://www.insilico.eu/coral>). In this approach, physicochemical properties and the experimental conditions of substance are represented by so-called quasi-SMILES, which are character-based representations derived from traditional Simplified Molecular Input Line Entry System (SMILES). Thus, a main advantage of the quasi-SMILES is to increase the number of available datasets by using the eclectic data in developing quasi-SMILES-based QSPRs/QSARs models. This chapter provides instructions on how to use CORAL software for building QSPR/QSAR models based on quasi-SMILES.

Keywords QSPR · QSAR · Eclectic information · Quasi-SMILES · CORAL software

S. Ahmadi (✉)

Department of Chemistry, Faculty of Pharmaceutical Chemistry, Tehran Medical Sciences, Islamic Azad University, Tehran, Iran

e-mail: s.ahmadi@iautmu.ac.ir; ahmadi.chemometrics@gmail.com

N. Azimi

Advanced Chemical Engineering Research Center, Razi University, Kermanshah, Iran

Abbreviations

AD	Applicability Domain
CCC	Concordance Correlation Coefficient
CORAL	CORrelation And Logic
CII	Correlation Intensity Index
EP	Endpoint
<i>F</i>	Fischer ratio
IIC	Index of Ideality Correlation
MAE	Mean Absolute Error
NPs	Nanoparticles
OECD	Organization of Economic Co-operation and Development
QSAR	Quantitative Structure–Activity Relationship
QSPR	Quantitative Structure–Property Relationship
RMSE	Root-Mean-Square Error
SMILES	Simplified Molecular Input Line Entry System
TF	Target Function

8.1 Introduction

Quantitative structure–activity/property relationship (QSAR/QSPR) approach is indubitably of considerable importance in food chemistry [1, 2], environmental chemistry [3], modern chemistry [4–6], biochemistry [7], nanotechnology [8, 9], and drug design [10, 11]. The QSAR/QSPR approach is the mathematical and computerized search for compounds with desired activities/properties using chemical intuition and experience. Once a structure–activity/property correlation has been established, any number of compounds, including those not yet synthesized, can be easily screened on a computer to select structures with the desired activity/properties. Then the most promising compounds can be found for synthesis and experimental testing [12]. Therefore, QSAR/QSPR study saves cost and time for the development process of new molecules as drugs, materials, additives, or any other purpose. While finding successful structure–activity models is not an easy task, the recent increase in the number of papers in QSPR/QSAR research clearly indicates the rapid evolution in this area. To obtain a significant correlation, it is very important to use appropriate descriptors, whether they are theoretical, empirical, or derived from easily empirical properties of the constructs [12]. A group of descriptors shows simple molecular properties and therefore can give insight into the physicochemical nature of the activity/property under consideration.

Considering the growth of nanotechnology, modeling the properties or toxicity of nanoparticles (NPs) on living organisms is very important [13–15]. Although it is difficult to conduct toxicological experiments or obtain physical properties of NPs on a case-by-case basis, QSPR/QSAR is a computationally efficient technique because

it saves time, cost, and animal sacrifice. The first part of nano-QSPR/QSAR model implementation includes data collection (including descriptors and endpoints) and data processing. The dataset can be obtained from the literature, databases, experiments, or integrated multiple sources. Therefore, to construct nano-QSPR/QSAR models, it is important to identify a new set of descriptors that can accurately represent the properties of NPs as well as the experimental conditions.

During recent years, the Simplified Molecular Input Line Entry System (SMILES) and quasi-SMILES descriptors have been examined by some researchers for QSPR/QSAR modeling [16–19]. The SMILES can reveal molecular structures, and quasi-SMILES can represent molecular structure and physicochemical properties and exposure conditions [8, 20, 21]. SMILES of a molecule is based on a set of rules that allow a molecular structure to be represented as a sequence of atom and bond symbols, but quasi-SMILES imports the physicochemical properties and experimental conditions as a string of characters after SMILES symbol.

8.2 Principals of QSPR/QSAR Models

Although QSPR/QSAR modeling has been used for over five decades, many studies still do not follow the Organization of Economic Co-operation and Development (OECD) guidelines. Figure 8.1 summarizes the best practices for each step of QSPR/QSAR approach using models in peer reviewed literature. Dearden et al. have reported a detailed description of common errors in QSPR/QSAR research [22].

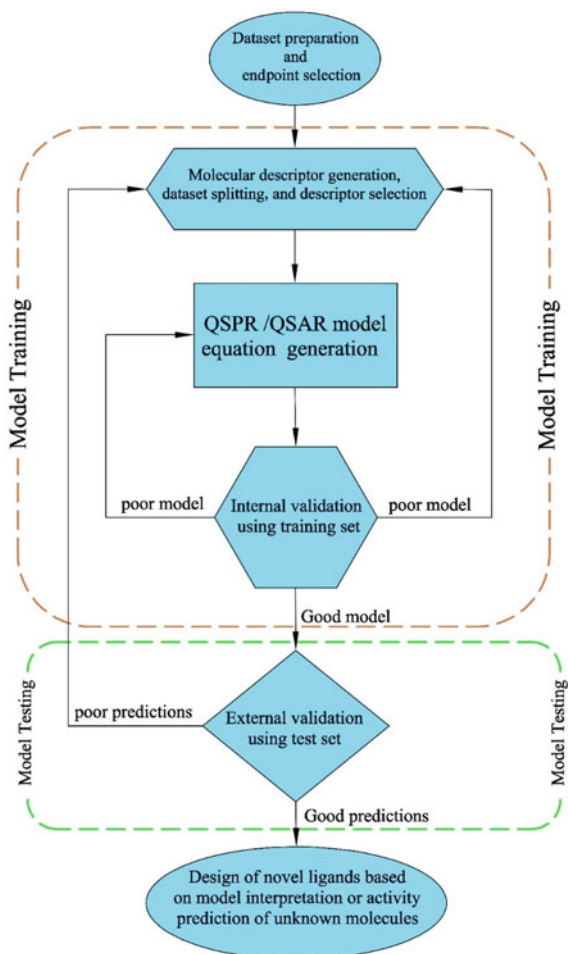
According to OECD guidelines, if a QSPR/QSAR study is to be reliable, the following five principles must be met: (i) a well-defined endpoint, (ii) an unambiguous algorithm, (iii) a defined applicability domain (AD), (iv) appropriate measures of goodness-of-fit, robustness, and predictivity, and (v) a mechanistic interpretation, if possible.

8.3 Monte Carlo Technique for Nano-QSPR/QSAR

8.3.1 SMILES and Quasi-SMILES

SMILES is a chemical notation system designed by Weininger et al. [23, 24]. According to the principles of molecular graph theory, SMILES uses a very small, natural grammar to specify precise structural features. The SMILES symbol system is also suitable for fast machine processing. Quasi-SMILES is an alternative to SMILES, which is used for substances considering physicochemical properties and experimental conditions.

Fig. 8.1 General flowchart for QSPR/QSAR modeling



8.3.2 *The Main Step for QSPR/QSAR Modeling by SMILES or Quasi-SMILES*

CORrelation And Logic (CORAL) software (<http://www.insilico.eu/coral>) has two possibilities for building QSPR/QSAR models based on SMILES or quasi-SMILES. In the following, the method of preparing the input data for the CORAL software is described.

(a)				(b)			
Set	ID	SMILES	EC ₅₀	Set	ID	Quasi-SMILES	Bandgap
+	1	<chem>c1{[N+](=O)[O-]}ccc(cc1)N</chem>	3.51	-	3	O=[Al]O[Al]=OA28	3.1
+	2	<chem>c1(ccccc1)C#N</chem>	2.90	+	4	O=[Al]O[Al]=OBN5	4.6
+	3	<chem>c1(Nc2c(ccc2)Cl)nc(nc(n1)Cl)Cl</chem>	5.19	+	5	O=[Al]O[Al]=OBN4	4.2
#	4	<chem>c1(Oc2ccccc2)ccc(cc1)Br</chem>	5.93	-	7	O=[Al]O[Al]=OA21	5.8
-	5	<chem>c1(Oc2ccc(cc2)Cl)ccc(cc1)N</chem>	5.55	+	10	O=[Al]O[Al]=OIZ4	4.06
+	6	<chem>C1(NC2CCCC2)CCCCC1</chem>	4.52	#	11	O=[Al]O[Al]=OIZ4	4
#	7	<chem>c1(Oc2ccccc2)ccccc1</chem>	5.14	#	12	O=[Al]O[Al]=OJZ4	4.2
#	8	<chem>N(CCCC)(CCCC)CCO</chem>	3.29	-	13	O=[Al]O[Al]=OJZ6	4.15
-	9	<chem>N(CCCC)(CCCC)CCCC</chem>	3.47	-	14	O=[Al]O[Al]=OJZ7	4.09
*	10	<chem>C(CCl)CCl</chem>	4.14	-	15	O=[Al]O[Al]=OIN1	3.88
#	11	<chem>c1(CNC(C)C)ccccc1</chem>	4.53	#	16	O=[Al]O[Al]=OAN1	3.6
#	12	<chem>c1(/N=N/c2ccccc2)ccccc1</chem>	5.04	-	17	O=[Ce]=OBN1	3.44
+	13	<chem>C1(CO1)CCC=C</chem>	3.37	*	18	O=[Ce]=OBN1	3.38
*	14	<chem>c1(/C=C/COC(=O)C)ccccc1</chem>	4.13	+	19	O=[Ce]=OEQ2	2.78
+	15	<chem>c1(ccccc1)N=C=S</chem>	5.64	+	20	O=[Ce]=OAR1	3.33
+	16	<chem>c1(ccccc1)NC(=O)C</chem>	2.66	+	21	O=[Ce]=OAN1	3.49
+	17	<chem>c1(ccc(cc1)OC)CCC(=O)C</chem>	3.64	*	22	O=[Ce]=OAN1	3.38
*	18	<chem>c1(ccc(cc1)N)CCCCCCCCC</chem>	6.48	#	23	O=[Ce]=OKN1	3.03

Fig. 8.2 Sample of data based on a SMILES, and b quasi-SMILES as input for CORAL

8.3.2.1 Dataset Preparation for Models Based on SMILES

The SMILES string is a procedure for representing a two-dimensional molecular graph as a one-dimensional string that can show the connectivity and chirality of a molecule. In most cases, there are too many SMILES strings for a structure. Canonical SMILES gives a single ‘canonical’ form for any particular molecule. Molecular structures of desired compounds were transformed to canonical SMILES using different software such as Open Babel and ACD/ChemSketch program. Figure 8.2a, b indicates the sample of data based on SMILES, and quasi-SMILES as input for CORAL software, respectively. The first column indicates set, the second is compound ID, the third is SMILES/quasi-SMILES, and the last column is desired property/activity.

8.3.2.2 Dataset Preparation for Models Based on Quasi-SMILES

For building of QSPR/QSAR in different physicochemical properties and/or the experimental conditions of substance, one can use quasi-SMILES instead of SMILES of molecules. Dataset preparation for quasi-SMILES is same as SMILES, only SMILES is replaced by quasi-SMILES.

8.3.2.3 Quasi-SMILES Definition for Various Datasets/Endpoints

Quasi-SMILES is a sequence of symbols that not only represents the molecular structure but also the different conditions that can affect the endpoint under investigation. Eclectic data can include: different physical properties such as temperature,

Table 8.1 Distinction of standardized physicochemical features into classes 1–9 according to its value

Normalized value	Class
$\text{Norm}(E) > 0.9$	9
$0.8 < \text{Norm}(E) < 0.9$	8
$0.7 < \text{Norm}(E) < 0.8$	7
$0.7 < \text{Norm}(E) < 0.6$	6
$0.6 < \text{Norm}(E) < 0.5$	5
$0.5 < \text{Norm}(E) < 0.4$	4
$0.4 < \text{Norm}(E) < 0.3$	3
$0.3 < \text{Norm}(E) < 0.2$	2
$0.2 < \text{Norm}(E) < 0.1$	1
$\text{Norm}(E) < 0.1$	0

pressure, and assay of experiment to obtain an endpoint, or cell line type, time exposition, concentration, etc. to obtain an activity. The type and number of eclectic data can be different in various datasets.

Quasi-SMILES may be made by eclectic condition, only [4, 13] or combination of SMILES and eclectic conditions [5, 8]. The continuous eclectic conditions can be normalized by the following equation for assigning codes:

$$\text{Norm}(E_i) = \frac{\min(E_i) + E_i}{\min(E_i) + \max(E_i)} \quad (8.1)$$

E_i is its value of physicochemical parameter E , $\min(E_i)$ is minimum value of E , and $\max(E_i)$ indicates maximum value of E .

According to Table 8.1, the number of unique values in each parameter was less than 10; therefore, the quasi-SMILES descriptors representations could be coded by assigning a number between zero and nine in a single character.

A further development of the CORAL software (CORAL-2020) allows the display of experimental conditions through groups of symbols enclosed in parentheses. Table 8.2 shows the comparison codes in the last version (CORAL-2020) and old version of CORAL for creating quasi-SMILES in recently proposed models for cytotoxicity of metal oxide NPs [4]. One can see codes-2020 are quite transparent and consequently are more convenient for a user. As is clearly evident, CORAL-2020 codes being quite transparent and thus more user-friendly. Table 8.2 indicates codes used for the cell line, method, time exposition, concentration, nanoparticle size, and metal oxide type. Table 8.3 indicates the examples of quasi-SMILES obtained based on these codes.

Toropov and Toropova developed a QSAR model based on the new version of CORAL for the toxicity of ZnO NPs [14]. Experimental data from the literature are toxicity assessment of ZnO NPs and ZnO NPs coated with polyethylene glycol (PEG), which are investigated by intraperitoneal injections in the rat (50, 100, 200 mg/kg) for one month. Measurement of the toxic effects of renal factors

Table 8.2 Codes used for the cell line, method, time exposition, concentration, nanoparticle size, and metal oxide type to convert various information of the experimental data to quasi-SMILES [4]

Feature	Value or type	Code	Code 2020	Feature	Value or type	Code	Code 2020
Cell line	MCF-7	H	[MCF-7]	Normalized NPs size	$0.2 < \text{Norm}(\text{size}) \leq 0.3$	P	$[0.2 < \text{Norm}(\text{size}) \leq 0.3]$
	HT-1080	I	[HT-1080]		$0.3 < \text{Norm}(\text{size}) \leq 0.4$	Q	$[0.3 < \text{Norm}(\text{size}) \leq 0.4]$
	HepG-2	J	[HepG-2]		$0.4 < \text{Norm}(\text{size}) \leq 0.5$	R	$[0.4 < \text{Norm}(\text{size}) \leq 0.5]$
	HT-29	K	[HT-29]		$0.5 < \text{Norm}(\text{size}) \leq 0.6$	S	$[0.5 < \text{Norm}(\text{size}) \leq 0.6]$
	PC-12	L	[PC-12]		$0.9 < \text{Norm}(\text{size}) \leq 1.0$	T	$[0.9 < \text{Norm}(\text{size}) \leq 1.0]$
Method	MTT	M	[MTT]	Metal oxide type	SnO ₂	1	[SnO ₂]
	NRU	N	[NRU]		MnO ₂	2	[MnO ₂]
Time exposition	24	X	[T24]		ZnO	3	[ZnO]
	48	Y	[T48]		Bi ₂ O ₃	4	[Bi ₂ O ₃]
	72	Z	[T72]		NiO	5	[NiO]
Concentration ($\mu\text{g mL}^{-1}$)	5	A	[C5]		CeO ₂	6	[CeO ₂]
	10	B	[C10]		SiO ₂	7	[SiO ₂]
	25	C	[C25]		TiO ₂	8	[TiO ₂]
	50	D	[C50]				
	100	E	[C100]				
	200	F	[C200]				

Table 8.3 Some examples for quasi-SMILES extracted by codes indicated in Table 8.2

Cell line	Method	Time exposition (h)	Concentration ($\mu\text{g mL}^{-1}$)	Normalized NPs size	Metal oxide type	Quasi-SMILES	Quasi-SMILES (2020)	Cell viability (%)
MCF-7	MTT	24	5	$0.2 < \text{Norm}(\text{size}) < 0.3$	SnO ₂	HMXAP1	[MTT][T24] [C5][0.2 < Norm(size) < 0.3][SnO ₂]	97.0
MCF-7	MTT	24	25	$0.2 < \text{Norm}(\text{size}) < 0.3$	MnO ₂	HMXAP2	[MTT][T24] [C25][0.2 < Norm(size) < 0.3][MnO ₂]	81.0
HT-1080	NRU	24	10	$0.9 < \text{Norm}(\text{size}) < 0.1$	MnO ₂	INXBP2	[NRU][T24] [C10][0.9 < Norm(size) < 0.1][MnO ₂]	94.0
MCF-7	MTT	48	100	$0.2 < \text{Norm}(\text{size}) < 0.3$	ZnO	HMYEP3	[MTT][T48] [C100][0.2 < Norm(size) < 0.3][ZnO]	4.1
HepG2	MTT	72	200	$0.2 < \text{Norm}(\text{size}) < 0.3$	SiO ₂	JMZFP7	[MTT][T72] [C200][0.2 < Norm(size) < 0.3][SiO ₂]	57.0
HepG2	NRU	72	5	$0.2 < \text{Norm}(\text{size}) < 0.3$	SiO ₂	JNZAP7	[NRU][T72] [C5][0.2 < Norm(size) < 0.3][SiO ₂]	95.7
PC12	MTT	48	50	$0.3 < \text{Norm}(\text{size}) < 0.4$	TiO ₂	LMYDR8	[MTT][T48] [C50][0.3 < Norm(size) < 0.4][TiO ₂]	59.0
HepG2	MTT	24	5	$0.5 < \text{Norm}(\text{size}) < 0.6$	NiO	JMXAS5	[MTT][T24] [C5][0.5 < Norm(size) < 0.6][NiO]	99.5
HT-29	MTT	24	50	$0.3 < \text{Norm}(\text{size}) < 0.4$	CeO ₂	KMXDQ6	[MTT][T24] [C50][0.3 < Norm(size) < 0.4][CeO ₂]	91.0
MCF-7	MTT	24	200	$0.9 < \text{Norm}(\text{size}) < 1.0$	Bi ₂ O ₃	HMXFT4	[MTT][T24] [C200][0.9 < Norm(size) < 1.0][Bi ₂ O ₃]	51.7

Table 8.4 Codes used as fragments of quasi-SMILES and their meaning

Code	Meaning
[15d]	Renal factor measured after fifteen days post-injection
[30d]	Renal factor measured after thirty days post-injection
[RF1]	Variation in creatinine as renal factor
[RF2]	Variation in uric acid as renal factor
[RF3]	Variation in blood urea nitrogen as renal factor
[50]	50 mg per kg of body weight
[100]	100 mg per kg of body weight
[200]	200 mg per kg of body weight
[ZnO]	Uncoated ZnO NPs is injected
[ZnO][peg]	ZnO coated by PEG NPs is injected

including creatinine, uric acid, and blood urea nitrogen was measured after 15 and 30 days after injection. Table 8.4 shows the quasi-SMILES attributes together with experimental conditions. Table 8.5 represents examples of available quasi-SMILES obtained based on this condition and related activity.

Toropova et al. developed new nano-QSAR model for predicting toxicity of nano-mixtures to *Daphnia magna* based on quasi-SMILES [25]. The binary mixtures of TiO₂ NPs and with of one of the second component including AgNO₃, Cd(NO₃)₂, Cu(NO₃)₂, CuSO₄, Na₂HAsO₄, NaAsO₂, benzylparaben, and benzophenone-3 have been investigated. Quasi-SMILES contain the following information: (1) Second

Table 8.5 Some examples for quasi-SMILES extracted by codes presented in Table 8.4

Time exposition (days)	Renal factor type	NPs (mg/kg)	NPs type	Quasi-SMILES	Experimental renal factor
15	Creatinine	50	ZnO	[15d][RF1][50][ZnO]	0.79
15	Creatinine	100	ZnO	[15d][RF1][100][ZnO]	0.87
15	Creatinine	100	ZnO-peg	[15d][RF1][100][ZnO][peg]	0.50
15	Uric acid	100	ZnO-peg	[15d][RF2][100][ZnO][peg]	1.37
15	Blood urea nitrogen	100	ZnO-peg	[15d][RF3][100][ZnO][peg]	62.30
30	Creatinine	100	ZnO	[30d][RF1][100][ZnO]	0.72
30	Uric acid	50	ZnO-peg	[30d][RF2][50][ZnO][peg]	1.30
30	Blood urea nitrogen	50	ZnO-peg	[30d][RF3][50][ZnO][peg]	50.33
30	Blood urea nitrogen	200	ZnO-peg	[30d][RF3][200][ZnO][peg]	49.0

Mixed substance	Core diameter of TiO ₂ NPs	Zeta potential of TiO ₂ NPs	Mole fraction of TiO ₂ NPs	Mol fraction of mixed substance	Exposure time (h)
	C	Z	F1	F2	E
Cd(NO ₃) ₂	30	-16.3	0.948	0.052	48
AgNO ₃	20	-1.8	0.999	0.001	48
AgNO ₃	20	-1.8	0.999	0.001	7
Na ₂ HAsO ₄	20	-1.8	0.982	0.018	48
Na ₂ HAsO ₄	20	-1.8	0.980	0.020	48

Experimental data

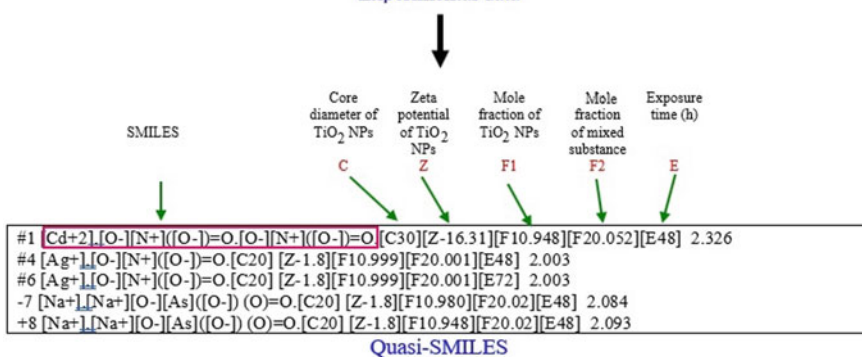


Fig. 8.3 Transfer of experimental data into quasi-SMILES [25]

component of mixture represented by SMILES; (2) core diameter of TiO₂ NPs; (3) Zeta potential of TiO₂ NPs; (4) mole fraction of TiO₂ NPs; (5) mole fraction of mixed substance; and (6) exposure time. Figure 8.3 shows the transformation of the experimental condition and substance into the quasi-SMILES.

8.3.2.4 Model Development

Model development has several steps that can be organized in CORAL software and does not require any software for data partitioning, descriptor generation, and model validation. In the following sections, the main step for QSPR/QSAR modeling using CORAL software is described.

8.3.2.5 Dataset Splitting

After the preparation and curation of dataset, the next step of building a QSAR/QSPR model for an endpoint by CORAL software (<http://www.insilico.eu/coral>) is loading an array of lines. Each line consists of four components.

The first column is the types of set which '+', '-', '#', and '*' indicate the active training, passive training, calibration, and validation, respectively (Fig. 8.2).

- The second column without space with type of set is number or ID of compound.
- The third column is quasi-SMILES.
- The last column is endpoint value.

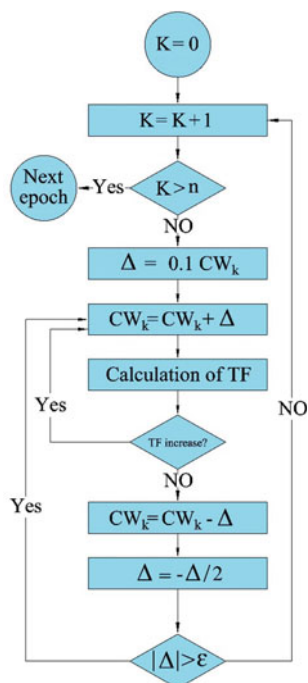
After the preparation of input file, the dataset was splitted into training, passive training, calibration, and validation sets using CORAL software, randomly with desired present for each set.

8.3.2.6 Monte Carlo Optimization Process

Quasi-SMILES is a group of attributes where each attribute group is converted into a group of coefficients called correlation weights. Monte Carlo optimization refines the correlation weights that provide numerical data on them, which maximizes the predictive potential of a model as much as possible. Figure 8.4 shows the flowchart of one cycle of Monte Carlo optimization of correlation weights (n is the number of correlation weights that contribute to model construction).

There are different target functions (TFs) in CORAL software for Monte Carlo optimization [25–29], which are introduced below four TFs:

Fig. 8.4 Flowchart of one cycle of the Monte Carlo optimization for finding correct correlation weights (n is the number of correlation weights that contribute to model construction)



$$TF_0 = r_{AT} + r_{PT} - |r_{AT} - r_{PT}| \times C \quad (8.2)$$

$$TF_1 = TF_0 + IIC_C \times W_{IIC} \quad (8.3)$$

$$TF_2 = TF_1 + CII_C \times W_{CII} \quad (8.4)$$

$$TF_3 = TF_2 + IIC_C \times W_{IIC} + CII_C \times W_{CII} \quad (8.5)$$

r_{AT} and r_{PT} represent the correlation coefficient between the experimental and predicted endpoints for active and passive training sets, respectively. Empirical constant (C), W_{IIC} , and W_{CII} have a defined numerical value [1, 18, 30–33].

IIC_C is the index of ideality correlation. IIC_C is obtained based on the calibration set as follows:

$$CII_C = r_c \frac{\min(-MAE_C, +MAE_C)}{\max(-MAE_C, +MAE_C)} \quad (8.6)$$

$$-MAE_C = \frac{1}{-N} \sum |\Delta_i|, \quad -N \text{ is the number of } \Delta_i < 0 \quad (8.7)$$

$$+MAE_C = \frac{1}{-N} \sum |\Delta_i|, \quad +N \text{ is the number of } \Delta_i \geq 0 \quad (8.8)$$

$$\Delta_i = \text{Obs}_i - \text{Calc}_i \quad (8.9)$$

The Obs_i and Calc_i are the experimental and predicted endpoint for i th compound.

The correlation intensity index (CII), like IIC criteria, was developed to modify the quality of the Monte Carlo optimization used to build the QSPR/QSAR models. CII is formulated as follows:

$$CII = 1 - \sum \Delta R_i^2 > 0, \text{ If } \Delta R_i^2 < 0 \text{ then } \Delta R_i^2 = 0 \quad (8.10)$$

$$\Delta R_i^2 = R_i^2 - R^2 \quad (8.11)$$

where R^2 is the coefficient of determination for all endpoints and R_i^2 is the coefficient of determination for all endpoints in the absence of i th compound. Therefore, if ΔR_i^2 is greater than zero, the meaning of i th is an ‘opposite’ for the correlation between the experimental and calculated values of the set.

A small sum of ΔR_i^2 means a more ‘intensive’ correlation.

The CORAL model for an endpoint (EP) is defined by the below equation:

$$EP = C_0 + C_1 \times DW(T, N) \quad (8.12)$$

C_0 and C_1 represent regression coefficients, T is a threshold, and N is the number of optimization cycles. The DCW(T, N) is defined as the below equation:

$$\text{DCW}(T, N) = \sum \text{CW}(S_k) \quad (8.13)$$

where S_k represents the symbol of a quasi-SMILES line; the $\text{CW}(S_k)$ shows the correlation weights of S_k .

8.3.2.7 Applicability Domain

The AD of QSAR/QSAR models for CORAL software is determined in two steps based on the distribution of SMILES or quasi-SMILES features in the training and calibration sets:

Step 1: the statistical defect (d_k) is calculated for each involved (unblocked) SMILES or quasi-SMILES feature (S_k) to build the model with the following equation:

$$d_k = \frac{|P(S_k) - P'(S_k)|}{N(S_k) + N'(S_k)} \quad (8.14)$$

here, $P(S_k)$ and $P'(S_k)$ represent the probability of S_k in the active training set and calibration sets, respectively; $N(S_k)$ and $N'(S_k)$ denote the frequencies of S_k in the active training and calibration sets, respectively.

Step 2: the quasi-SMILES (D_i) statistical defect of all compounds is defined according to the following equation:

$$D_i = \sum_{k=1}^{N_A} d_k \quad (8.15)$$

here N_A denotes the number of non-blocked quasi-SMILES features in the quasi-SMILES.

Quasi-SMILES falls in the AD if:

$$D_i < 2 \times \overline{D} \quad (8.16)$$

where \overline{D} represents average statistical defect of the training set.

8.3.2.8 Model Validation

Validation, as the fourth principle of OECD, is recognized as an intrinsic component to check the robustness, predictability, and reliability of any QSPR/QSAR models. There are three approaches to examine the robustness, reliability, and predictive potential of the QSPR/QSAR models in CORAL software, including:

- Internal validation
- External validation
- Y-scrambling or data randomization.

Various statistical criteria such as determination coefficient (R^2), concordance correlation coefficient (CCC), cross-validated correlation coefficient (Q^2), Q_{F1}^2 , Q_{F2}^2 , Q_{F3}^2 , standard error of estimation (s), mean absolute error (MAE), Fischer ratio (F) and root-mean-square error (RMSE), R_m^2 , and average of R_m^2 metric ($\overline{R_m^2}$) are calculated to authenticate the QSPR/QSAR models constructed based on the Monte Carlo optimization by the CORAL software. Table 8.6 indicates the mathematical equation of diverse statistical benchmark of the predictive potential for CORAL models.

Table 8.6 Mathematical formulation of different statistical benchmark of the predictive potential for CORAL models

Criterion of the predictive potential	Description	References
$Q^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2}$	Leave-one-out cross-validated correlation coefficient	[34]
$Q_{F1}^2 = 1 - \frac{\sum_{i=1}^{N_{EXT}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N_{EXT}} (\hat{y}_i - \bar{y}_{TR})^2}$	Criteria of predictability	[35]
$Q_{F2}^2 = 1 - \frac{\sum_{i=1}^{N_{EXT}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N_{EXT}} (\hat{y}_i - \bar{y}_{EXT})^2}$	Criteria of predictability	[35]
$Q_{F3}^2 = 1 - \frac{[\sum_{i=1}^{N_{EXT}} (\hat{y}_i - y_i)^2] / N_{EXT}}{[\sum_{i=1}^{N_{EXT}} (\hat{y}_i - \bar{y}_{EXT})^2] / N_{TR}}$	Criteria of predictability	[36]
$R_m^2 = R^2 \times \left(1 - \sqrt{R^2 - R_0^2}\right)$		[36]
$\overline{R_m^2} = \frac{R_m^2(x,y) - R_m^2(y,x)}{2}$	Average of R_m^2 metric	[36]
$CCC = \frac{2 \sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2 + \sum(y - \bar{y})^2 + n(\bar{x} - \bar{y})^2}$	Concordance correlation coefficient	[37]
$C_{R_p^2} = R \sqrt{R^2 - R_f^2}$	Coefficient of determination for Y-randomization	[38]

8.3.2.9 Mechanistic Interpretation

The 5th OECD principle focuses on mechanistic interpretation of the QSPR/QSAR model if possible. The model interpretation is used to examine the critical and responsible attributes that influence the endpoint. Finally, the new compounds are designed based on these attributes. In the QSPR/QSAR modeling based on the CORAL software, the same structural attributes (S_k) collected from three or more different splits are used to perform the mechanistic interpretation [39–42]. These structural attributes (S_k) are divided into three categories according to previous studies:

- Increasing factor if the $CW(S_k)$ is positive in all splits and in three attempts,
- Decreasing factor if the $CW(S_k)$ is negative in all splits and in three attempts,
- Undefined attributes if the $CW(S_k)$ is both positive and negative [43–45].

8.4 Examples of Quasi-SMILES-Based QSPR/QSAR Models

Some examples of QSAR/QSPR models base on quasi-SMILES with CORAL software using different TFs are presented in Table 8.7.

8.5 Conclusion and Future Direction

QSPR/QSAR modeling based on SMILES and quasi-SMILES by CORAL software is useful for big dataset. In CORAL software, QSPR/QSAR generally follows the five OECD principles. In addition, additional principles may be defined practically for nano-QSPR/QSAR that reflect the nature of the nanomaterial under investigation. For example, the new principles should take into account the test conditions and the quality of the applied equipment.

The use of CORAL software in building QSPR/QSAR models for nanomaterials in different conditions is simple, and the models can be easily predicted and interpreted. There are very good TFs (TF_0 – TF_3) to find reliable correlation weights and this is one of the important capabilities of CORAL for building excellent QSAR/QSAR models. The type and number of input features can change the performance of a QSAR/QSPR model. But there is one of a shortcoming for CORAL software, the user can use only CORAL software descriptors, and it is impossible to add the other descriptors produced by other descriptor generators.

In CORAL software, there is only Monte Carlo algorithm to find correlation weights. The use of various algorithms can increase the quasi-SMILES QSPR/QSAR performance. Data splitting in CORAL software is done randomly; the possibility of using different methods of data splitting can increase the validity of the models.

Table 8.7 Some examples of QSAR/QSPR models base on quasi-SMILES with CORAL software using different TFs

Compound	Endpoint	No. of quasi-SMILES	Eclectic data										
			Second component of mixture	Core diameter of TiO ₂ NPs	Zeta potential of TiO ₂ NPs		Mole fraction of TiO ₂ NPs	Mole fraction of NPs	Mole fraction of TiO ₂	Mole fraction of substance			
Nano-mixtures	EC50 for <i>Daphnia magna</i>	67	Second component of mixture	Core diameter of TiO ₂ NPs	Zeta potential of TiO ₂ NPs		Mole fraction of TiO ₂ NPs	Mole fraction of NPs	Mole fraction of TiO ₂	Mole fraction of substance	Exposure time	Exposure time	Mole fraction of substance
Nano-mixtures	EC50 for <i>Daphnia magna</i>	67	Second component of mixture	Core diameter of TiO ₂ NPs	Zeta potential of TiO ₂ NPs		Mole fraction of TiO ₂ NPs	Mole fraction of NPs	Mole fraction of TiO ₂	Mole fraction of substance	Exposure time	Exposure time	Mole fraction of substance
Ag NPs	pL _{C50} for <i>Daphnia magna</i> and zebrafish	170	Status of NPs (bare, coat, cons)	Core diameter of TiO ₂ NPs	Organisms (<i>Daphnia</i> or zebrafish)		Mole fraction of TiO ₂ NPs	Mole fraction of NPs	Mole fraction of TiO ₂	Mole fraction of substance	Exposure time	Exposure time	Mole fraction of substance
Metal oxide NPs	Cell viability (%)	83	Cell line	Assay	Time exposition	Concentration	NP size	NP size	Metal oxide type	Metal oxide type	–	–	–
Metal–organic frameworks	Log(CO ₂ uptake)	260	BET	Specific surface area	Pore volume	Pressure	Temperature	Temperature	–	–	–	–	–
Metal oxide NPs	Band gap	198	Synthesis method	–	Annealing temperature	–	Crystalline size	Crystalline size	–	–	–	–	–
Cadmium containing quantum dots	Hepatic cell viability (%)	115	Core	Norm diameter	Charge	Surface modification	Assay type	Assay type	Exposure time	Delivery type	Exposure time	Exposure time	QD concentration
CdSe quantum dots with ZnS shell	HeLa cell viability (%)	61	Size	Surface ligand	Surface charge	Surface modification	Surface modification	Surface modification	Assay type	Exposure time	Exposure time	Exposure time	QD concentration

(continued)

Table 8.7 (continued)

Compound	Endpoint	No. of quasi-SMILES	Eclectic data						
			Cell line	Assay	Time exposition	Concentration	NPs size	Metal oxide type	
Metal oxide NPs	Cell viability (%)	83	Cell line	Assay	Time exposition	Concentration	NPs size	Metal oxide type	–
Multiwalled carbon nanotubes	Cell viability (%) of human lung cells	255	Mean diameter	Mean length	Surface area	Toxic assay method	Cell line	Exposure time	Dose
Metal oxide NPs	Cell viability (%) of human lung and skin cells	336	Core size	Hydrodynamic size	Surface charge	Dose	Toxic assay method	Cell line	–
Nanofluids	Viscosity ratio	100	Size	Volume fraction (%)	–	–	–	–	–
Nanofluids	Viscosity ratio	100	Size	Volume fraction (%)	–	–	–	–	–
Nanofluids	Viscosity ratio	100	Shape	Volume fraction (%)	–	–	–	–	–
Nanofluids	Viscosity ratio	100	Shape	Volume fraction (%)	–	–	–	–	–
Nanozeolites	Cell viability (%)	120	Time exposition	Concentration	Cell type	Sample	–	–	–
Metal oxide NPs	Cell membrane damage	137	Chemical element	Concentration	Time of exposure	–	–	–	–

(continued)

Table 8.7 (continued)

Compound	TF	Statistical indicator value				References
		R^2_{Train}	R^2_{Val}	S_{Train}	S_{Val}	
Nano-mixtures	TF ₁	0.64–0.90	0.32–0.80	0.34–0.72	0.48–0.95	[25]
Nano-mixtures	TF ₂					
Ag NPs	TF ₃		0.62–0.71	0.34–0.55	0.58–0.60	[46]
Metal oxide NPs	TF ₁	0.92–0.93	0.87–0.94	7.8–9.0	8.6–13.1	[4]
Metal–organic frameworks	TF ₁	0.74–0.77	0.74–0.77	0.25–0.29	0.20–0.26	[5]
Metal oxide NPs	TF ₁	0.85–0.88	0.82–0.90	0.33–0.38	0.32–0.38	[8]
Cadmium containing quantum dots	TF ₁	0.70–0.90	0.63–0.81	8.7–14.5	–	[9]
CdSe quantum dots with ZnS shell	TF ₁	0.84–0.96	0.59–0.93	0.11–0.35	0.27–0.51	[3]
Metal oxide NPs	TF ₂	0.95–0.97	0.87–0.94	5.21–6.53	6.2–11.0	[13]
Multiwalled carbon nanotubes	TF ₀	0.60–0.80	0.81–0.88	13.7–15.2	8.0–16.4	[47]
Metal oxide NPs	TF ₀	0.71–0.73	0.70–0.76	–	–	[20]
Nanofluids	TF ₁	0.84–0.92	0.73–0.90	0.75–0.11	0.05–0.08	[18]
Nanofluids	TF ₂	0.85–0.91	0.90–0.94	0.08–0.10	0.04–0.09	
Nanofluids	TF ₁	0.74–0.88	0.82–0.92	0.25–0.57	0.26–0.42	
Nanofluids	TF ₂	0.73–0.85	0.90–0.94	0.40–0.51	0.20–0.51	
Nanozeolites	TF ₀	0.83–0.89	0.72–0.81	0.02–0.04	0.02–0.03	[48]
Metal oxide NPs	TF ₀	0.50–0.54	0.67–0.93	0.38–0.39	0.25–0.40	[49]

Since the correlation weight of the descriptors in this software is calculated through Monte Carlo approach, the use of consensus modeling can dramatically increase the prediction results.

References

1. Ahmadi S, Ghanbari H, Lotfi S, Azimi N (2021) *Mol Divers* 25(1):87–97. <https://doi.org/10.1007/s11030-019-10026-9>
2. Achary PGR, Toropova AP, Toropov AA (2019) *Food Res Int* 122:40–46. <https://doi.org/10.1016/j.foodres.2019.03.067>
3. Kumar A, Kumar P (2021) *J Hazard Mater* 402:123777. <https://doi.org/10.1016/j.jhazmat.2020.123777>
4. Ahmadi S (2020) *Chemosphere* 242:125192. <https://doi.org/10.1016/j.chemosphere.2019.125192>
5. Ahmadi S, Ketabi S, Qomi M (2022) *New J Chem* 46:8827–8837. <https://doi.org/10.1039/D2NJ00596D>
6. Lotfi S, Ahmadi S, Kumar P (2021) *RSC Adv* 11:33849–33857. <https://doi.org/10.1039/D1RA06861J>
7. Ahmadi S, Khazaei MR, Abdolmaleki A (2014) *Med Chem Res* 23:1148–1161. <https://doi.org/10.1007/s00044-013-0716-z>
8. Ahmadi S, Aghabeygi S, Farahmandjou M, Azimi N (2021) *Struct Chem* 32:1893–1905. <https://doi.org/10.1007/s11224-021-01748-4>
9. Kumar P, Kumar A (2021) *Nanotoxicology* 15:1199–1214. <https://doi.org/10.1080/17435390.2021.2008039>
10. Ghasedi N, Ahmadi S, Ketabi S, Almasirad A (2022) *J Recept Signal Transduct* 42:418–428. <https://doi.org/10.1080/10799893.2021.1988971>
11. Ahmadi S, Moradi Z, Kumar A, Almasirad A (2022) *J Recept Signal Transduct* 42:361–372. <https://doi.org/10.1080/10799893.2021.1957932>
12. Karelson M, Lobanov VS, Katritzky AR (1996) *Chem Rev* 96:1027–1044. <https://doi.org/10.1021/cr950202r>
13. Ahmadi S, Toropova AP, Toropov AA (2020) *Nanotoxicology* 14:1118–1126. <https://doi.org/10.1080/17435390.2020.1808252>
14. Toropov AA, Toropova AP (2021) *Sci Total Environ* 772:145532. <https://doi.org/10.1016/j.scitotenv.2021.145532>
15. Toropov AA, Toropova AP (2020) *Sci Total Environ* 737:139720. <https://doi.org/10.1016/j.scitotenv.2020.139720>
16. Ahmadi S, Akbari A (2018) *Environ Res* 29:895–909. <https://doi.org/10.1080/1062936X.2018.1526821>
17. Lotfi S, Ahmadi S, Kumar P (2022) *RSC Adv* 12:24988–24997. <https://doi.org/10.1039/D2RA03936B>
18. Jafari K, Fatemi MH, Toropova AP, Toropov AA (2022) *Chemom Intell Lab Syst* 222:104500. <https://doi.org/10.1016/j.chemolab.2022.104500>
19. Toropov A, Toropova A, Lombardo A, Roncaglioni A, Lavado G, Benfenati E (2021) *Environ Res* 32:463–471. <https://doi.org/10.1080/1062936X.2021.1914156>
20. Choi J-S, Trinh TX, Yoon T-H, Kim J, Byun H-G (2019) *Chemosphere* 217:243–249. <https://doi.org/10.1016/j.chemosphere.2018.11.014>
21. Lotfi S, Ahmadi S, Zohrabi P (2020) *Struct Chem* 31:2257–2270. <https://doi.org/10.1007/s11224-020-01568-y>
22. Dearden JC, Cronin MTD, Kaiser KLE (2009) *Environ Res* 20:241–266. <https://doi.org/10.1080/10629360902949567>

23. Weininger D (1988) *J Chem Inf Model* 28:31–36. <https://doi.org/10.1021/ci00057a005>
24. Weininger D, Weininger A, Weininger JL (1989) *J Chem Inf Comput Sci* 29:97–101. <https://doi.org/10.1021/ci00062a008>
25. Toropova AP, Toropov AA, Fjodorova N (2022) *NanoImpact* 28:100427. <https://doi.org/10.1016/j.impact.2022.100427>
26. Kumar P, Kumar A, Lal S, Singh D, Lotfi S, Ahmadi S (2022) *J Mol Struct* 1265:133437. <https://doi.org/10.1016/j.molstruc.2022.133437>
27. Azimi A, Ahmadi S, Kumar A, Qomi M, Almasirad A (2022) *Polycycl Aromat Compd* 1–21. <https://doi.org/10.1080/10406638.2022.2067194>
28. Ahmadi S, Lotfi S, Afshari S, Kumar P, Ghasemi E (2021) *Environ Res* 32:1013–1031. <https://doi.org/10.1080/1062936X.2021.2003429>
29. Ahmadi S, Mehrahi M, Rezaei S, Mardafkan N (2019) *J Mol Struct* 1191:165–174. <https://doi.org/10.1016/j.molstruc.2019.04.103>
30. Nimbhal M, Bagri K, Kumar P, Kumar A (2020) *Struct Chem* 31:831–839. <https://doi.org/10.1007/s11224-019-01468-w>
31. Toropova AP, Duchowicz PR, Saavedra LM, Castro EA, Toropov AA (2020) *Mol Inform* 39:1900070. <https://doi.org/10.1002/minf.201900070>
32. Toropova AP, Toropov AA, Carneseccchi E, Benfenati E, Dorne JL (2020) *Environ Sci Pollut Res* 27:13339–13347. <https://doi.org/10.1007/s11356-020-07820-6>
33. Kumar P, Kumar A (2021) *J Mol Struct* 1246:131205. <https://doi.org/10.1016/j.molstruc.2021.131205>
34. Shayanfar A, Shayanfar S (2014) *Eur J Pharm Sci* 59:31–35. <https://doi.org/10.1016/j.ejps.2014.03.007>
35. Consonni V, Ballabio D, Todeschini R (2009) *J Chem Inf Model* 49:1669–1678. <https://doi.org/10.1021/ci9000115y>
36. Roy K, Kar S (2014) *Eur J Pharm Sci* 62:111–114. <https://doi.org/10.1016/j.ejps.2014.05.019>
37. Lin LI-K (1992) *Biometrics* 48:599. <https://doi.org/10.2307/2532314>
38. Rucker C, Rucker G, Meringer M (2007) *J Chem Inf Model* 47:2345–2357. <https://doi.org/10.1021/ci700157b>
39. Manisha, Chauhan S, Kumar P, Kumar A (2019) *Environ Res* 30:145–159. <https://doi.org/10.1080/1062936X.2019.1568299>
40. Kumar P, Kumar A, Sindhu J, Lal S (2019) *Drug Res (Stuttg)* 69:159–167. <https://doi.org/10.1055/a-0652-5290>
41. Kumar P, Kumar A, Sindhu J (2019) *Environ Res* 30:63–80. <https://doi.org/10.1080/1062936X.2018.1564067>
42. Kumar P, Kumar A, Sindhu J (2019) *SAR QSAR Environ Res* 30:525–541. <https://doi.org/10.1080/1062936X.2019.1629998>
43. Toropov AA, Toropova AP, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2012) *Anticancer Agents Med Chem* 12:807–817. <https://doi.org/10.2174/187152012802650255>
44. Nesmerak K, Toropov AA, Toropova AP, Kohoutova P, Waisser K (2013) *Eur J Med Chem* 67:111–114. <https://doi.org/10.1016/j.ejmech.2013.05.031>
45. Veselinović AM, Milosavljević JB, Toropov AA, Nikolić GM (2013) *Eur J Pharm Sci* 48:532–541. <https://doi.org/10.1016/j.ejps.2012.12.021>
46. Toropov AA, Kjeldsen F, Toropova AP (2022) *Chemosphere* 303:135086. <https://doi.org/10.1016/j.chemosphere.2022.135086>
47. Trinh TX, Choi J-S, Jeon H, Byun H-G, Kim J (2018) *Chem Res Toxicol* 31:183–190. <https://doi.org/10.1021/acs.chemrestox.7b00303>
48. Leone C, Bertuzzi EE, Toropova AP, Toropov AA, Benfenati E (2018) *Chemosphere* 210:52–56. <https://doi.org/10.1016/j.chemosphere.2018.06.161>
49. Toropova AP, Toropov AA, Benfenati E, Korenstein R, Leszczynska D, Leszczynski J (2015) *Environ Sci Pollut Res* 22:745–757. <https://doi.org/10.1007/s11356-014-3566-4>

Chapter 9

Quasi-SMILES-Based Mathematical Model for the Prediction of Percolation Threshold for Conductive Polymer Composites



Swayam Aryam Behera, Alla P. Toropova, Andrey A. Toropov,
and P. Ganga Raju Achary

Abstract The traditional method for creating conductive polymer composites (CPCs) involves mixing carbon black, metal powder, or carbon fibre into a polymer matrix. Since the polymer matrix acts as an insulator, when a threshold filler level is achieved, the conductivity of these composites can exhibit a sharp increase. The common term generally used to describe such phenomena is called ‘percolation’. As the conductive filler content increases in the insulator polymer matrix, it creates different conductive routes, steady rise in the electrical conductivity is observed at a critical volume fraction Φ . That critical volume fraction Φ responsible for the transition of polymers from insulators to conducting is called the ‘*percolation threshold*’. The diverse experimental percolation threshold cured data of 45 conductive polymer composite systems were classified into four sets: A = active training set; P = passive training set; C = calibration set; V = validation set. Systems of eclectic conditions of various processes of mixing such as dry mixing, latex technology, and melt blending employed to fabricate the conducting polymer composites with various polymer matrixes like high-density polyethylene (HDPE), low-density polyethylene (LDPE), maleic anhydride (MA), polyamide (PA) and the conducting fillers such as multi-wall carbon nanotube (MWNT), single-wall carbon nanotube (SWNT), polyaniline (PANI) are very important and crucial to have desired properties. Unique quasi-SMILES codes for different CPCs were suggested taking into consideration various systems of eclectic conditions. These quasi-SMILES codes were the basis for building mathematical models for predicting percolation threshold CPCs.

S. A. Behera · P. G. R. Achary (✉)

Department of Chemistry, Institute of Technical Education and Research (ITER), Siksha ‘O’ Anusandhan University, Bhubaneswar, Odisha 751030, India

e-mail: pgrachary@soa.ac.in

A. P. Toropova · A. A. Toropov

Laboratory of Environmental Chemistry and Toxicology, Department of Environmental Health Science, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156 Milano, Italy

Keywords Conductive polymer composites · Percolation threshold · Quantitative structure–property relationship (QSPR) · Quasi-SMILES

9.1 Introduction

Electrically conductive polymer composites (CPCs), which are made of metallic, carbonaceous or conducting polymeric particles dispersed in a multi-phase blend or a single polymer matrix, have drawn considerable industry and academic attention over several decades [1–5]. The number of research publications about CPCs that were found on 20th May 2014 when the term ‘conductive polymer composite’ was searched in the (Web of Science database) Institute for Scientific Information (ISI), serves as evidence of their popularity. CPCs have served applications as electromagnetic interference (EMI) shielding, conductors, and sensors due to their low cost, ease of processing, and tunable electrical characteristic compared to intrinsic conducting polymers [6–9]. The specific applications of CPCs depend on their electrical resistivity (Table 9.1). For instance, EMI shielding necessitates electrical resistivity values of $10^{-2}\Omega\text{ cm}$, whereas CPC materials for electrostatic dissipation typically require an electrical resistivity of $10^{-6}\Omega\text{ cm}$ in plastic fuel tanks.

The CPCs having electrical performance depend only on conductive (continuous) networks built after inserting the conductive fillers because the majority of common host polymers are fundamentally insulating [10, 11]. The CPC material will demonstrate an insulator/conductor transition at a critical level when the conductive filler content reaches; particularly, the electrical conductivity dramatically increases when the initial conducting channels are produced by several orders of magnitude. The percolation threshold Φ_c is referred to as this critical volume fraction Φ . As the conductive filler content rises, additional conductive routes may be created in the

Table 9.1 Classifying conductive polymer composite materials according to their electrical resistivity and application ranges

Resistivity ($\Omega\text{ cm}$)	Applications and products
Insulating (10^{11} to 10^{14})	Insulators
Electrostatic dissipative (10^6 to 10^{11})	Anti-static materials: microscope housing materials, fuel tanks, anti-static storage containers, electronic connectors, electrostatic paintable compounds, mining pipes, etc.
Conductive (10^1 to 10^6)	Sensors & EMI shielding: electronic nose devices, strain sensing materials, self-regulated heating elements, organic liquid sensing devices, over-current protectors, etc.
Highly conductive (10^{-6} to 10^1)	Conductors: conducting adhesives & coatings. Resistors, bipolar plates, metal replacement, bus bars, thermos-electric materials, etc.

polymer matrix, steadily increasing the electrical conductivity until a plateau is achieved at its saturation. Typically, a power law can be used for a CPC material to objectively describe the electrically conductive behaviour [10].

$$\sigma = \sigma_0(\Phi - \Phi_c)^t \quad (9.1)$$

where σ is the electrical conductivity of the CPC and t is the critical exponent for the conductive networks related to the dimensionality in the CPC. For two-dimensional (2D) conductive networks, this model uses $t \approx 2$, and for three-dimensional (3D), $t \approx 1.3$. However, the experimental values frequently differ from these expected values [12, 13].

The melt-mixing technologies, such as internal mixing, twin-screw extrusion, and injection moulding, are the widely used approaches among the conventional CPC fabrication methods (i.e. melt mixing, solution processing, and in situ polymerisation) used to fabricate commercial CPC materials. This is because current industrial practices are compatible with these techniques. However, traditional melt-mixing techniques typically have a high Φ_c in which the CPCs are made. Theoretically, the 16 vol% percolation value anticipated by the classical percolation theory [14, 15] is close to 10–20 vol% of the Φ_c for randomly dispersed, spherical fillers, such as metallic particles, carbon black (CB), and conducting polymer particles.

Although carbon nano-tubes (CNT) and graphene nano-sheets (GNS) have huge surface areas that can sustain well-developed transport networks, these high-aspect-ratio conductive nanoparticles' severe agglomeration characteristic during host polymers processing produces the high Φ_c relatively. Unfortunately, a number of disadvantages are in CPCs with high Φ_c , including low economic viability, high-melt viscosities, and worse mechanical qualities, particularly in terms of ductility and toughness [16, 17]. Therefore, high-performance CPC materials manufacturing, lowering Φ_c efficiently has emerged as a persistent, significant concern.

The most promising method for achieving low Φ_c [18–20] in a CPC material has remained the formation of a segregated structure. Throughout the entire CPC system, instead of being distributed randomly conductive fillers are largely found at the polymeric matrix particle interfaces in segregated CPC (s-CPC) materials. Several times this particular structure reduce the percolation value as compared to ordinary melt-mixed CPCs because in the interfacial regions of s-CPC materials, there is perfect mutual contact and an extremely high percentage between the conductive fillers. For example, in acrylonitrile–butadiene–styrene (ABS), Gupta et al. created a segregated CB-based conductive network with an exceptionally 0.0054 vol% of low Φ_c value, the lowest value for CPC materials (CB-based) in the literature at this time [21]. A polymeric matrix with conductive fillers and an exclusionary microstructure assigned a constrained volume is the basis for the formation of a segregated conductive network mechanism, which at specific filler concentrations, the effective density of the conductive pathways significantly raises. In a nutshell, with little filler loading, this intriguing topology offers an effective paradigm for establishing a conductive network.

As depicted in Fig. 9.1, there have been three primary methods developed to prepare s-CPCs. To create segregated conductive networks, the first method is using dry or solution mixing, a combination of polymer granules compression coated with conductive fillers (Fig. 9.1a) [22, 23]. Different conductive fillers (e.g. CB, metallic particles, GNSs, and CNTs) on the external surfaces of polymeric particles can be distributed without overly emphasising the filler dispersion levels before being hot compressed to form bulk materials with segregated structures due to the simplicity of the mixing and compaction processing methods [24–27]. However, due to processing difficulties, the filler concentration cannot reach very high values (often less than 10 wt%), and the polymers should have relatively high-melt viscosities utilised with this construction approach to sustain the segregated conducting networks during hot compression moulding. The second method, known as latex technology, involves spreading conductive fillers into polymeric latex. The fillers are kept between the latex particles within the interstitial spaces while the polymer emulsion is freeze-dried (Fig. 9.1b) [18, 28, 29]. In spite of the somewhat sophisticated manufacturing technology, this method has clear advantages: when compared to materials made through dry or solution mixing, latex materials made using only distilled water have the following advantages: (i) an environmentally friendly and inexpensive process; (ii) a satisfactory dispersion at the surfaces of the latex particles of conductive fillers; (iii) and the availability of any composition of polymer-filler systems without being constrained by high-melt viscosities during melt-mixing [30, 31].

The third tactic relies on melt blending, which is at the interfaces of immiscible polymer mixes and conductive fillers' selective distribution (Fig. 9.1c) [32, 33]. Melt blending is the initial option when producing s-CPC products industrially because of how straightforward it is. However, because this method encompasses so many influencing factors, such as kinetics parameters (such as sequence and mixing procedures, shear strength, and blending time), thermodynamic coefficients (such as the interfacial energy between the conductive fillers and polymer matrices), and forming a stable segregated conductive network are significantly more challenging than it is for other technologies at the interfaces of polymer blends [34–36]. The 'segregated conductive network concept' for nickel particle/(HDPE) high-density polyethylene composites was first put forth by Turner and colleagues in 1971 [22, 23]. Since then, s-CPCs based on conductive fillers and various polymeric matrices have undergone extensive research to determine the relationships between processing, morphology, and property. In order to maximise their performance, the associated CPC variables (such as grain size, polymer modulus, and processing and parameter) have also been identified.

The present chapter highlights the importance of the percolation threshold, the effect of conductive filler and host polymers, different methods generally employed to fabricate the conductive polymer, the changes in the conducting properties, applications of such conductive polymer systems and a theoretical attempt to build a mathematical model to predict percolation threshold for conductive polymer composites.

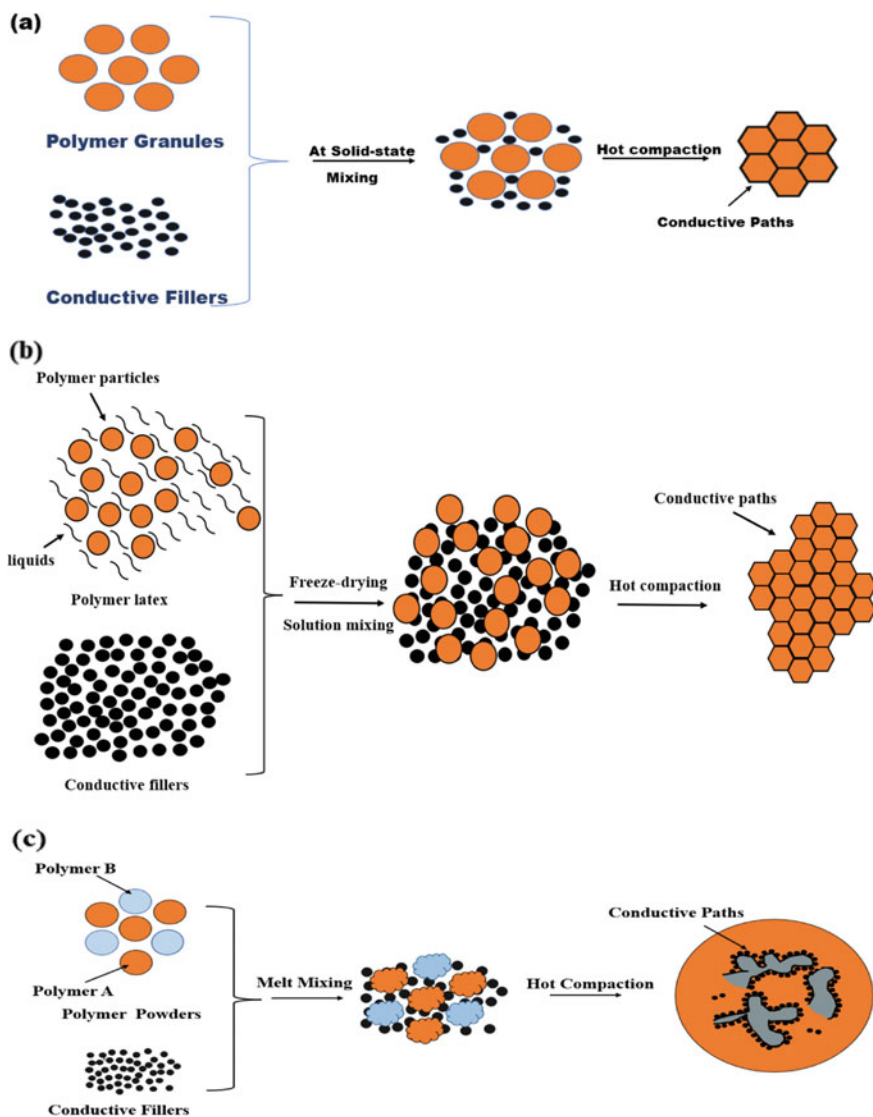


Fig. 9.1 Schematic for the fabrication of the s-CPCs using various processing methods: **a** dry or solvent mixing, **b** latex technology, and **c** melt-blending methods

9.2 Theoretical Background of the Percolation Threshold

Significant effort has been put towards customising separated structures during the past ten years to achieve ultralow Φ_c . The ultralow percolation behaviours, such as conductive filler type, the polymeric matrix, and fabrication procedures, help explain

the highly intriguing electrical percolation behaviours of s-CPCs. Polymers having high-melt viscosity, like polystyrene (PS), ultrahigh molecular weight polyethylene (UHMWPE), and natural rubber (NR), make up the majority of s-CPC matrices because they can preserve the conductive pathways that are confined in the interfacial areas during processing. Furthermore, conductive fillers with high aspect ratios (such GNSs and CNTs) have drawn more attention than those with low aspect ratios (e.g. metallic particles, CB, and graphite flakes). Great-aspect-ratio fillers are quite popular because of their excellent transport characteristics and high effectiveness in constructing segregated conductive networks.

The s-CPC systems at very low (below 0.5 vol%) conductive filler loadings typically change from being an insulator to a conductor, as seen in row 4 of Table 9.1. The values for UHMWPE-based s-CPCs varied from 0.028 to 0.5 vol%, depending on the unique morphology of segregated conductive networks and the types of conductive fillers. Despite this, there does not appear to be agreement regarding the Φ_c of s-CPCs. After inserting large polymeric beads for the segregated conductive networks (about a diameter of 5 mm) as a scaffold, Gerhardt's group in a segregated CB/ABS system was able to achieve the lowest (0.0054 vol%) Φ_c recorded among s-CPC materials [21].

The CPCs with emulsion-based and melt-blended have greater Φ_c than those made through solution mixing or dry technology when the impact of the dispersion methods on Φ_c of s-CPCs is examined. By latex technology, the relatively high Φ_c of s-CPCs produced can be attributed to two factors: (i) the size of the polymeric latex particles (typically at the nanometre level) is too small to achieve even distribution with the conductive fillers [37, 38], and (ii) the low melt viscosity of the latex polymer makes the conductive fillers more difficult to easily stabilise at the interface between the polymeric matrix granules. The conductive fillers are entrenched in the polymeric matrix during mixing in the s-CPCs created by melt compounding, which reduces their effectiveness in creating segregated conducting networks [31, 39].

The range of values for σ_{\max} , another critical s-CPC parameter, is wide (10^{-7} to 10^4 S/cm). The obvious discrepancies in the maximum values in Table 9.1 may be explained by the junction resistance between the conductive fillers and the inherent electrical conductivity of the conductive fillers. For instance, the transport characteristics of thermally or chemically reduced GNSs are poorer; hence, the s-CPCs, which are based on CNT, always demonstrate greater conductance than the GNS-based materials [24, 40–42]. High junction resistance is caused by the segregated conductive channels, which are caused by host polymer layers forming insulating gaps between the neighbouring conductive fillers or surfactants like (SDS) sodium dodecyl sulphate [43, 44]. Additionally, the inter-diffusion of the molecular chains is primarily prevented by the segregated distribution of conductive fillers, which hinders the melting process, especially at high loading levels [19].

As a result, the filler weight percentage of s-CPCs made using dry, melt-blending, and mechanical processes cannot be greater than 10%. The melt viscosity has little bearing on the filler concentration of the s-CPCs manufactured using the latex method, which can range practically between 0 and 100 wt% anywhere [2]. As a result, the σ_{\max} of the s-CPCs made using dry, melt-blending, and mechanical

procedures is often smaller than that of the materials manufactured using latex technology. Several works [45–47] must be acknowledged in order to expand the σ_{\max} number of these CPC. Instead of the usual insulating stabilisers, they used an intrinsically conducting polymer system, specifically poly(3,4-ethylenedioxythiophene): poly(styrene sulfonate) (PEDOT: PSS), to disperse the CNTs as the conducting surfactant. At 30 vol% CNTs, they achieved high σ_{\max} values (10^3 S/cm), which is comparable to the electrical conductivity of pristine CNTs. Theoretically, thanks to the segregated distribution, conductive fillers can build a typical two-dimensional conductive network [24, 48, 49]. The intricacy of the segregated conductive networks and many contributing elements, such as the morphology, dispersion, and distribution of conductive fillers, are attributed to this phenomenon [30]. This is accomplished by attributing the variations in the percolation behaviours of the s-CPC materials to a variety of factors, such as the electrical properties, processing techniques and dispersion quality of the conducting fillers, and the modulus, molecular weight, and particle size of the polymeric matrices.

9.2.1 *Effect of the Conductive Fillers*

The sort of conductive fillers significantly impacts how electrically conductive s-CPC materials are. In this section, we focus on the impact of geometrical morphology, aspect ratio, intrinsic electrical conductivity, and dispersion techniques for conductive fillers on the c , \max , and t of s-CPC systems. According to the excluded volume theory [50], the Φ_c of the s-CPCs decrease as the conductive filler aspect ratio is increased when the conductive fillers are uniformly dispersed at the interfaces between polymeric domains. Grossiord et al. discovered that the high-aspect-ratio (~ 120) MWNTs reduced to 20% of Φ_c that for the low-aspect-ratio MWNTs (~ 40) for the PS-based s-CPCs made by the latex technique [51]. The percolation behaviours of PVAc-based s-CPCs filled with low-aspect-ratio CB and high-aspect-ratio SWNTs were explored by Grunlan et al. [18, 28, 52]; the Φ_c of the SWNT/PVAc s-CPCs reached an ultralow value (0.03 vol%), which is significantly lower than that of the CB ones (2.39 vol%). These two instances show that creating segregated conductive networks is frequently made easier by high-aspect-ratio conductive fillers. Additionally, the conductive fillers with high aspect ratios always result in greater maximum values. According to Mierczynska and colleagues, the UHMWPE-based s-CPC materials were less conductive than the MWNT ones due to the high level of SWNT agglomeration [53]. Therefore, while building segmented conductive networks, effective dispersion techniques are required to achieve the benefits of high-aspect-ratio conductive fillers.

The majority of the conductive filler in the separated CPCs remains at the polymer domains' interface. As a result, it is challenging to maintain the uniform dispersion of the conductive fillers, particularly the high-aspect-ratio ones at relatively high loadings. The MWNTs continue to localise as aggregates at the surfaces of the UHMWPE

granules despite the extensive sonication and mechanical stirring treatment. High-aspect-ratio conductive nano-fillers are not advised for use in the construction of segregated conductive networks without effective dispersion methods due to their low economic affordability and efficiency (e.g. adding surfactants followed by intensive sonication and mechanical stirring measurements).

The geometry of the conductive fillers inseparably affects the t of the segregated conductive networks because t largely depends on the distribution of the tunnelling distance within the CPC materials [12, 13]. Using the two most common high-aspect-ratio fillers, 2D GNS, and one-dimensional (1D) CNT, as examples, the 2D GNS segregated networks frequently show a value of t below 1.3. Disparities in the conductive network microstructure may be the reason for the variable dimensions of the segregated conductive networks. Due to the flat nature of 2D GNSs, the nano-sheets are frequently restacked, resulting in a segregated conductive network formed by plane-to-plane contact, which accounts for the low dimensionality of the segregated conductive networks. Furthermore, compared to 1D CNTs, 2D structure fillers, GNSs were less inclined to interlace and build high-dimensional conductive networks [24].

Due to changes in the intrinsic electrical conductivity and dispersion of the conductive fillers, their chemical surface qualities also have an impact on the dimensionality of the segregated conductive networks [54]. In order to stabilise the placement of GNSs at the interfaces between the PS and PMMA phases, Tan et al. functionalised GNSs covalently with P(St-co-MMA) recently. This s-CPC displayed an odd value of t that reached a maximum of 6.9 and was indicative of a complete departure from the conventional percolation hypothesis [54]. Due to the insulating layers of grafted molecular chains coated on the conductive nano-sheets and the poor electrical conductivity of the chemically modified GNS, t underwent a significant divergence. In the following section, we'll talk about how conductive fillers affect the σ_{\max} s-CPC.

The σ_{\max} of s-CPCs is determined by the junction resistance between nearby conductive fillers and the inherent electrical conductivity of those fillers. The intrinsic electrical conductivity of impurities, such as amorphous carbon, catalyst particles, and surface imperfections, in traditional carbon fillers is typically below 10^2 S/cm. Due to its comparatively low electrical conductivity, high σ_{\max} values are less accessible, even at large loadings, and conductor applications, such as those for bipolar plates, conducting polymer adhesives, and thermoelectric materials, are not possible [44]. However, it appears that metallic fillers with better intrinsic electrical conductivities would be better candidates for achieving reasonably high σ_{\max} [55–57].

9.2.2 *Effect of the Host Polymers*

The chemical and physical characteristics of host polymers, which serve as scaffolds for segregated conductive networks, inexorably impact the electrical performance

of S-CPCs. The host polymeric matrices' molecular weight and modulus affect the percolation behaviours. Because there is less mixing of the host polymers and conductive fillers during hot compaction, host polymers with large molecular weights and moduli are better able to resist plastic deformation [25, 36]. The s-CPCs that have high molecular weights and moduli in their polymer matrix consequently invariably have relatively low Φ_c . In contrast to low molecular weight (1×10^6 g/mol) one, the Φ_c (~ 1.0 wt%) of the CB/UHMWPE s-CPCs with a high molecular weight (6×10^6 g/mol) showed attenuation of about 100%. The σ_{\max} increased to 10^{-2} S/cm for the high molecular weight material from 10^{-4} S/cm for the low molecular weight UHMWPE s-CPCs [25].

9.3 Methods for the Synthesis of Conductive Polymers

9.3.1 Chemical Method

Conductive polymers (CPs) have been created chemically by polymerising matching monomers after they have undergone oxidation or reduction. The potential for affordable mass production is one of its benefits. To improve the yield and quality of the manufactured product produced using the oxidative polymerisation process, numerical studies have been used. The employment of electrochemical techniques is not mandated by chemical route principles [58]. For instance, the well-known and widely researched CP poly(3-hexylthiophene) is virtually always created chemically. Chemical methods can be used to manufacture polypyrrole (PPy) and polyaniline (PANI); however, electrochemical methods typically result in variations with higher conductivity and mechanical qualities. After conjugation, stability is the primary requirement when getting ready for chemical polymerisation. Oligomers and low molecular weight polymers must be sufficiently reactive and soluble to polymerise in order for high molecular weight polymerisation to be successful. The polymerisation should continue using a heterogeneous technique if an oligomer precipitates out of the solution, although this is becoming less and less likely as the concentration of monomer and reactive polymer decreases. A failed chemical polymerisation would stop before the molecular entanglement weight is reached, leaving the reaction vessel walls with a mechanically unstable covering. However, chemical polymerisation guarantees the exact choice of oxidant to selectively create cation radicals at the appropriate position on the monomer in an adequately soluble system.

9.3.2 Metathesis Method

The interchange of one component from each substance to create a new one is known as metathesis, and it occurs when two chemicals interact chemically. Ring-opening cyclo-olefin metathesis, acyclic or cyclic alkynes metathesis, and di-olefin metathesis are the three types of metathesis polymerisation. Evans et al. investigated the metathesis of derivatives of aniline and 1,2-dihydroquinoline [59]. Masuda has examined the characteristics of polymers created using metathesis polymerisation that is typically based on acetylene [60].

9.3.3 Photochemical Method

The primary techniques for locating polymers in industry and academic research facilities have been chemical approaches [61]. However, during the past two decades, although extensively studied, photochemical preparation has been claimed to have minimal advantages due to its speed, low cost, and environmental friendliness. The technique can be used to fabricate some CPs. As an illustration, pyrrole has been successfully polymerised to PPy by exposure to visible light while acting as either a suitable electron acceptor or photosensitiser.

9.3.4 Electro-Chemical Method

Among the various described synthesis techniques, electrochemical synthesis of CPs is highly important since it is straightforward, affordable, can be carried out in a single-section glass cell, is reproducible, and the generated films have the necessary thickness and homogeneity. Anodic oxidation of suitable electroactive functional monomers is the electrochemical method utilised the most frequently to prepare electro-CPs; cathodic reduction is employed much less frequently. In the earlier example, the simultaneous creation of a polymer layer and the doping of counter ions as a result of oxidation takes place. The capacity for monomer oxidation leading to polymerisation is frequently higher than the potential for charging oligomeric intermediate polymers. A streamlined method of electropolymerisation, using alternate chemical and electrode reaction stages, was used to polymerise an electroactive monomer, such as pyrrole or thiophene [62]. For instance, in the potential dynamic electropolymerisation of thiophene, a radical cation is typically likely to form in the initial electrode reaction stage of thiophene electro-oxidation, cleared by an anodic peak of high positive potential [63]. At the subsequent chemical reaction stage, the radical cation reacts with the monomer to produce the protonated dimer of a radical cation. Then, during the electrode reaction step, the protonated dimer of the radical cation is electro-oxidised to the decomposition.

9.3.5 Plasma Polymerisation

An innovative method for creating thin films from a variety of organic and organometallic starting ingredients is plasma polymerisation. Pinhole-free and strongly cross-linked plasma polymerised films are insoluble, thermally stable, chemically inert, and physically robust. Furthermore, these films stick exceedingly well to various substrates, including those made of common polymer, glass, and metal surfaces [64]. They have been widely used in recent years for a variety of applications, including perm-selective membranes, protective shells, biological materials, electronic, optical devices, and adhesion supports, thanks to their exceptional qualities.

9.3.6 Solid-State Method

By using vacuum, heat, or removal with an inert gas to drive away reaction by-products, solid-state polymerisation enlarges polymer chain lengths in the absence of oxygen and water. Pressure, temperature, and the diffusion of waste products from the pellet's core to the shell all influence the reaction. After melt polymerisation, it is a crucial step frequently employed to improve polymers' mechanical and rheological characteristics before injection blow moulding [65]. This process is incredibly helpful in the commercial manufacturing PET films, advanced industrial fibres, and fibres suitable for bottles. The main industrial benefits of solid-state polymerisation are using straightforward, inexpensive equipment, and avoiding some of the issues with traditional polymerisation processes.

9.3.7 Inclusion Method

Atomic or molecular-level manufacturing of composite materials is often accomplished using inclusion polymerisation. Therefore, this type of polymerisation can open the door to extraordinary low-dimensional composite materials that have a lot of potentials. An electroconductive polymer, for instance, might be used to create a molecular wire. Composites of these polymers with organic hosts have been created based on inclusion. According to Miyata et al., this polymerisation can be seen as a typical space-dependent polymerisation and shouldn't only be seen from the standpoint of stereo-regular polymerisation [66]. The author failed to mention conventional solutions and bulk polymerisations in previous investigations.

9.4 Various Properties of Conducting Polymers

Conductive polymers are the subject of intensive research due to their exceptional qualities, such as tunable electrical properties, high optical, and mechanical capabilities. Conducting polymer composites have several uses in the electrical, electronic, and optoelectronic domains thanks to their synergistic effects.

9.4.1 Magnetic Properties

Due to their exceptional magnetic properties and technological implications, CPs' magnetism is greatly interesting. Transition metal oxide nanoparticles are crucial, in addition to the structural and magnetic properties of nanomaterials to be included in a polymer matrix. EPR and magnetisation measurements are the two basic experimental methods for examining the magnetic characteristics of conducting polymers [67]. EPR is highly sensible, and it is able to look at low energy changes in the produced polymers' magnetic characteristics that are related to unpaired electrons. On the other hand, magnetisation measurements track the samples' overall reaction to magnetic moments. Consequently, from this vantage point, these two methodologies offer complementing information.

9.4.2 Optical Properties

In optical absorption, i.e. in an excited state, a pi electron can be promoted from the lower energy state to the highest energy state in a tiny molecule with an isolated double bond by absorbing a photon with energy greater than the energy gap (E_g) between the two orbitals. However, a comparable molecule with conjugated double bonds will have an energy difference between its lowest unoccupied molecular orbital (LUMO) and its highest occupied molecular orbital (HOMO). A lower energy photon can encourage a pi electron from HOMO to LUMO because orbital interactions reduced the energy gap; as a result, in conducting polymers, the energy gap E_g can be even smaller [68].

However, in excited state relaxation through optical emission, a semiconducting polymer can boost an electron from HOMO to LUMO and create an exciton. This electron-hole pair is electrostatically bonded when the polymer absorbs a suitable energy photon. This excited state species can move from one place to another until it relaxes due to some deactivation process. Luminescence is one of the most practical methods for deactivating conducting polymers (light emission).

9.4.3 *Electrical Properties*

The doping level, chain arrangement, conjugation length, and sample purity all affect how the conductivity of polymers is determined. Electrical CPs lack long-range organisation and are molecular in origin. Electronic motion occurs around the individual macromolecules because polymers are molecular. For polymers and inorganic semiconductors, different processes are used to achieve high conductivity. The development of self-localised excitons such as solitons, polarons, and bi-polarons is related to the higher conductivities, which depend on doping in the polymers. These particles result from a powerful interaction between the charges on the chain that doping enabled. Charged solitons are the charge carriers in CPs with degenerate ground states, such as trans-polyacetylene, while polarons are typically formed on doping in CPs with non-degenerate ground states, such as PPy. After that, these polarons combine to create spinless bi-polarons, which are used as charge carriers [68]. The inexpensive cost of the polymers and the ability to molecularly design the appropriate characteristics have made them incredibly desirable materials for electrically conductive applications.

9.5 Applications of Conductive Polymers

9.5.1 *Sensors*

As an electrode modification, conductive polymers are used in sensor technologies to improve sensitivity, impart selectivity, minimise interference, and provide a support matrix for sensing materials. Below are some examples of sensors that use conductive polymers [69]:

- (a) **Gas Sensor:** A major ecological problem is the release of gaseous pollutants, including nitrogen oxide, SO₂, and hazardous gases from related businesses. To recognise and assess the concentration of such gaseous contaminants, sensors are necessary. Gas sensor equipment has typically been made using PANI and PPy.
- (b) **Humidity Sensor:** According to electrical, optical, and other physical properties, humidity sensors (HSs) are capable of detecting relative humidity in a variety of situations. The industrial and medical communities paid these sensors a lot of attention. Humidity calculations and regulation are important in various fields, including the food and electronics industries, residential environments, and medicine, among others. Humidity sensor devices have made use of the hydrophilic features of polymers, polymer composites, and modified polymers.
- (c) **Bio Sensor:** Conductive polymers are being employed in chemical analysis for the large-scale detection of ions and molecules in the liquid phase. Over the

past 20 years, the development of biosensors has been one of the most significant areas. It is reported that the most recent developments in biosensors and their applications are in the fields of agriculture, medicine, environmental monitoring, and clinical detection [70]. Conductive polymers might be used in the sensing mechanism or to immobilise the component that senses the molecular modifications. Biosensors have been created using the films produced by the electrochemical co-deposition of enzymes on CP or conductive substrates [71].

9.5.2 Solar Cells

Polymer solar cells (PSCs) have developed into a competitive substitute for silicon-based solar cells. PSCs provide a number of important benefits, including inexpensive production costs, straightforward processing, mechanical flexibility, and adaptability of a chemical structure due to advancements in organic chemistry. A plastic film substrate has been used in various experiments on flexible and lightweight appliances in place of fragile glass. A transparent anode must be applied using organic-based materials in order to create entirely plastic PSCs.

9.5.3 Supercapacitors

The popular name for a group of electrochemical capacitors is supercapacitors (SC). Because of their variety of uses, conductive polymers are a topic of interest to many researchers. Developing novel, specifically designed electrode materials with improved performance has received emphasis from SCs [72]. Conductive polymers, high-surface carbons, and transition metal oxides are typically used as SC electrode materials. Superior capacitive energy density and inexpensive cost of materials are two advantages of SCs based on CPs. Their main advantages are increased electrical conductivity, improved pseudo-capacitance, and a quick doping/de-doping rate during the charge/discharge process.

9.5.4 Data Storage Transistors

Due to their exceptional qualities, conductive polymers have found widespread use in electronics as charge storage and field effect transistors. Due to its capacity to enhance in-situ and gate-modulate channel conductance, conductive polymers can be used as field effect transistors to achieve high sensitivity.

9.5.5 Batteries

The first area where conductive polymers are sure to have a significant commercial influence is this one. The electrolyte provides a physical separation between the cathode and the anode and provides a source of cations and anions to balance the redox processes [73]. The electrodes allow for the accumulation of current and the diffusion of power.

9.6 Mathematical Models for the Prediction of Percolation Threshold

Thus, the electrically conductive properties of polymeric materials find numerous and varied applications. However, the experimental tuning of polymer systems for certain tasks is complex (requiring time for qualified performers) and expensive (purely economic factors, such as the cost of materials, energy, and labour remuneration). Under such circumstances, the attractiveness of computational methods becomes quite obvious as a tempting alternative to direct experiments. In other words, the development of computer technologies for the development of appropriate models becomes an important or even an integral part of technologies related to the electrically conductive properties of polymer systems.

9.6.1 Data and Building the Quasi-SMILES Codes

The diverse experimental percolation threshold data of 55 conductive polymer composite systems were obtained from the literature [74]. The above data were manually cured to remove the duplicity in the data. The refined data of 45 best diverse system data were chosen to build a quasi-simplified molecular input-line entry systems (quasi-SMILES) [75, 76]-based quantitative structure–property relationships (QSPR) mathematical model to predict the percolation threshold theoretically.

Systems of eclectic conditions of various processes of mixing such as dry mixing, latex technology, and melt blending employed to fabricate the conducting polymer composites with different polymer matrixes like high-density polyethylene (HDPE), low-density polyethylene (LDPE), maleic anhydride (MA), polyamide (PA) and the conducting fillers such as multi-wall carbon nanotube (MWNT), single-wall carbon nanotube (SWNT), and polyaniline (PANI) are very important and crucial to have desired properties.

Table 9.2 contains a list of symbols and groups of symbols (quasi-SMILES atoms, i.e. fragments of quasi-SMILES line, which cannot be examined separately) which

Table 9.2 Details of the quasi-SMILES codes employed

Polymer matrix	Quasi-SMILES code	Fillers	Quasi-SMILES code	Process of mixing	Quasi-SMILES code
ABS	A	CB	1	Dry mixing	w
HDPE	B	GRAPHITE	2	Latex technology	x
BA	C	SWNT	3	Melt blending	y
MMA	D	GNS	4	Solution mixing	z
AAEM	E	MWNT	5		
LDPE	F	EG	6		
NR	G	Al	7		
PA	H	ITO	8		
PC	I	CNT	9		
PE	J	Clay	U		
PS	K	PANI	V		
PP	L	CU	0		
PET	M	CUNW	I		
PMMA	N				
PPS	O				
PVC	P				
PVDF	Q				
SAN	R				
UHMWPE	S				
WPU	T				
PVAc	U				

are utilised to represent various conditions. These systems were randomly split into the training ($\approx 65\text{--}70\%$), calibration ($\approx 15\text{--}17\%$), and validation ($\approx 15\text{--}17\%$) sets.

Table 9.3 lists the final quasi-SMILES Codes of each polymer composite system with their experimental *percolation threshold*.

9.6.2 Optimal Descriptor

The correlation weights of the various components that can be included in the construction of a model are used to calculate the best descriptors. The circumstance that occurs the most frequently is when data on molecular structure features

Table 9.3 Quasi-SMILES code of polymer composites with their experimental percolation threshold and critical exponent (Eq. 9.1)

S. No.	Quasi-SMILES code	Φ_c	t
1	A1w-	0.0054	
2	A2w-	0.16	1.68
3	B4z-	0.95	1.08
4	F1w-	1.0	
5	G4x-	0.62	
6	I4x-	0.14	4.04
7	BK1y+	0.40	
8	BM1y-	3.80	
9	N7w-	10.0	2.0
10	N1w-	0.26	
11	N3x+	0.20	
12	N4x-	0.16	
13	N8w+	3.0	
14	LKA1y-	0.95	2.90
15	L9x+	0.30	
16	L4x-	0.03	1.69
17	Kx+	0.28	1.58
18	K1x+	1.50	
19	K9z+	0.05	
20	K3x+	0.40	
21	K4x-	0.20	
22	K4x+	0.60	
23	Kiz-	0.67	
24	KN4z-	0.02	6.92
25	U1x-	2.39	1.57
26	U1ux-	0.90	
27	U3x+	0.04	
28	Uvx+	0.60	4.6
29	P5w-	0.05	3.50
30	P0w-	5.0	2.9
31	P4x+	0.30	
32	Q4z-	0.11	1.10
33	Q5z+	0.07	
34	Q5w+	0.08	1.04
35	R5z+	0.03	2.15
36	S1w-	0.26	2.90
37	S1w+	0.50	

(continued)

Table 9.3 (continued)

S. No.	Quasi-SMILES code	Φ_c	t
38	S5w+	0.50	
39	S5z-	0.07	1.13
40	S3w+	0.14	2.0
41	S5w-	0.06	1.80
42	S4z-	0.06	1.54
43	S4w-	0.10	1.17
44	SN5z-	0.09	0.37
45	S45z+	0.10	
46	T1x-	0.23	1.20

are used to create a model for an endpoint that would see quantitative structure–property–activity relationships (QSPRs/QSARs) based on molecular graphs [77–79]. Simplified molecular input-line entry systems (SMILES), which can also be used to construct QSPR/QSAR, are an alternative to the molecular graph [80].

When SMILES/quasi-SMILES is employed as the foundation for QSPR or QSAR, an endpoint is viewed as a mathematical function of the SMILES/quasi-SMILES nomenclature, such as

$$\text{Endpoint} = F(\text{SMILES/quasi-SMILES}). \quad (9.2)$$

However, there are occasions when an endpoint is a mathematical function of not just a particular chemical molecular structure but also of its physicochemical (temperature, pressure), biochemical (toxicity and/or mutagenicity), and/or both circumstances [75, 76]. Instead of using conventional SMILES, which represent the molecular structure, in these situations, one might utilise quasi-SMILES, which are lines of symbols that reflect not only molecular structure but also physicochemical and/or biological parameters that can have an impact on an endpoint [75, 76].

The foundation of the theoretical mathematical model to forecast the percolation threshold is the one-variable correlations between descriptor of correlation weights (DCW) calculated with correlation weights of quasi-SMILES fragments [75, 76] and various experimental percolation threshold data.

The following formula is used to compute the ideal descriptor:

$$\text{DCW}(T, N) = \sum \text{CW}(S_k) + \sum \text{CW}(SS_k) + \sum \text{CW}(SSS_k) \quad (9.3)$$

where the S_k , SS_k , and SSS_k are pieces of a quasi-SMILES line that, respectively, contain one, two, and three quasi-SMILES ‘atoms’. The quasi-SMILES atom is a collection of symbols that cannot be studied individually since they collectively represent a specific situation [81]. In Table 9.2, the groups that are used to construct quasi-SMILES are depicted.

For example, symbols H, I, J, M, N, L, O, U, P and Q represent polyamide (PA), polycarbonate (PC), polyethylene (PE), poly(ethylene terephthalate) (PET), poly(methyl methacrylate) (PMMA), polypropylene (PP), poly(phenylenesulfide) (PPS), poly(vinyl acetate) (PVAc), poly(vinyl chloride) (PVC), poly(vinylidene fluoride) (PVDF), respectively (Table 9.2). The quasi-SMILES codes are assigned for different fillers such as CB:1; GRAPHITE:2; SWNT:3; GNS:4; MWNT:5 (Table 9.2). Similarly, processes of mixing to fabricate the conducting polymer composites such as dry mixing, latex technology, melt blending, and solution mixing were assigned the quasi-SMILES codes as ‘w’, ‘x’, ‘y’, and ‘z’, respectively.

The optimisation process employing the Monte Carlo approach is used to determine the correlation weights of all S_k , SS_k , and SSS_k , i.e. $CW(S_k)$, $CW(SS_k)$, and $CW(SSS_k)$ [80]. The procedure has two parameters: (i) the T , which is the threshold for classifying quasi-SMILES fragments into rare and non-rare categories (correlation weights of quasi-SMILES fragments that are rare, according to the selected T , have correlation weight equal to zero); and (ii) the N , which is the number of optimisation epochs.

The correlation coefficient between the endpoint and descriptor, computed using Eq. 9.3 for the training set, is the desired outcome of the optimisation approach. When the calibration set’s correlation coefficient reaches its maximum, the operation should be ended. If the process is continued past this point, the model will likely exhibit overtraining (i.e. excellent statistical quality for the training set but poor quality for the calibration and the validation set).

Since $T = T^*$ and $N = N^*$ yield the highest correlation coefficient for the calibration set, there is where the model should start. These T^* and N^* ought to be established using computational studies using T from a range of T_1, T_2, \dots, T_n and N from a range of N_1, N_2, \dots, N_m . With the correlation weights produced in the method just explained, one can use Eq. 9.4 to determine the best descriptor for each system with eclectic circumstances and then create a model using the systems in the training set.

$$\text{Percolation Threshold } (\Phi_c) = C_0 + C_1 \times \text{DCW}(T^*, N^*) \quad (9.4)$$

The generated model should have predictive capability after cross-checking against the calibration set to ensure sufficient statistical quality. The validation set serves as the final estimate of the predictive potential for Eq. 9.4 in the stated model-building process.

9.7 Results and Discussion

These quasi-SMILES-based mathematical models for three random splits into the training, calibration, and validation sets are presented in the following Eqs. (9.5)–(9.7):

$$\begin{aligned} \text{Percolation Threshold } (\Phi_c) &= -1.2561(\pm 0.1050) \\ &+ 0.4076(\pm 0.0629) * \text{DCW}(1, 15) \end{aligned} \quad (9.5)$$

$$\begin{aligned} \text{Percolation Threshold } (\Phi_c) &= -1.7077(\pm 0.1087) \\ &+ 0.7008(\pm 0.0501) * \text{DCW}(1, 15) \end{aligned} \quad (9.6)$$

$$\begin{aligned} \text{Percolation Threshold } (\Phi_c) &= -1.0869(\pm 0.0653) \\ &+ 0.6465(\pm 0.0420) * \text{DCW}(1, 15) \end{aligned} \quad (9.7)$$

The statistical characteristics of the quasi-SMILES-based model for the prediction of the *percolation threshold* (Φ_c) of different conductive polymer composites are summarised in Table 9.4.

Table 9.5 lists the percentage of identity for three random splits adopted in the present study.

The list of structural attributes (SA) and their correlation weights with the defect SA_k for the above three models are represented in Tables 9.6, 9.7, and 9.8, respectively.

Figure 9.2 shows the experimental and predicted *percolation threshold* (Φ_c) of the above three models.

These ranges of the statistical characteristics of models for the validation set for models based on the correlation weights of quasi-SMILES fragments are:

$$r^2 \in [0.5082, 0.5504], \text{ RMSE} \in [0.371, 0.532]$$

Thus, the suggested models' level is overage compared with models from work [82]. However, the model (calculated with Eq. 9.4) is built up by utilising conceptually other approaches. In addition, in fact, the suggested approach is checked up with three different splits into the training, calibration, and validation sets (Table 9.4). In other words, the approach is reliable.

Here the Monte Carlo method using the CORAL software (<http://www.insilico.eu/coral>) has been applied. But it should be taken into account, the range of problems involved in modern polymer science is exactly the same as the range of problems in the natural sciences as a whole. Hence, many other approaches QSPR/QSAR uses to analyse the polymer systems. There are both QSPR analysis [83–89] and QSAR analysis devoted to polymer systems [90–95]. Along with polymer electrical conductivity, stability [83–85], thermodynamic properties [86–88], and viscosity of polymers [89] are important modelling objectives. Of considerable interest are studies devoted to QSAR analysis of polymer systems, both natural [84] and transport-oriented polymeric substances introduced through membranes, which can be drug deliverers and means of reducing undesirable environmental consequences [92]. Quantum mechanical approaches to the study of polymer systems are gradually becoming on the same flow as traditional quantum mechanical analysis applied to organic and inorganic

Table 9.4 Statistical characteristics of the quasi-SMILES-based model for the prediction of the percolation threshold (Φ_c)

	Set ^a	<i>n</i>	<i>R</i> ²	CCC	IIC	CHI	Q ²	Q ² _{F1}	Q ² _{F2}	Q ² _{F3}	RMSE	MAE	F
1	A	11	0.3014	0.4632	0.4575	0.7290	0.0394				0.735	0.644	4
	P	12	0.7973	0.1107	0.8929	0.8558	0.6702				0.965	0.880	39
	C	11	0.5795	0.6650	0.7599	0.7262	0.4237	0.8622	0.3771	0.9339	0.210	0.179	12
	V	12	0.5504								0.406	0.246	
2	A	12	0.5799	0.7341	0.7615	0.7369	0.4338				0.528	0.447	14
	P	11	0.5828	0.1971	0.4115	0.7349	0.0858				1.19	1.08	13
	C	11	0.4872	0.6469	0.6913	0.6028	0.1689	0.8878	0	0.8945	0.274	0.187	9
	V	12	0.5082								0.411	0.360	
3	A	11	0.5128	0.6779	0.4092	0.7684	0.3650				0.532	0.428	9
	P	11	0.6296	0.3931	0.7821	0.7807	0.0611				0.782	0.620	15
	C	12	0.6279	0.2631	0.7907	0.7594	0.4022	0	0	0.2387	0.710	0.529	17
	V	12	0.5433								0.371	0.294	

^a A = active training set; P = passive training set; C = calibration set; V = validation set

Table 9.5 Percentage of identity for three splits acceptable

$(i, j)^a$	Split 1	Split 2	Split 3
Split 1	0	43.5	45.5
Split 2	27.3	0	34.8
Split 3	27.3	54.5	0

^a Matrix Element $[i, j]$, $i > j$ the identity for the active training sets; Matrix Element $[i, j]$, $i < j$ the identity for the validation sets

Table 9.6 List of structural attributes (SA) and their correlation weights (CW) for the model (Eq. 9.5)

SA _k	CW(SA _k)	ID	N1	N2	N3	DEFECT[SA _k]
+...	0.3842	1	4	6	2	0.053
-...	0.7312	2	7	6	9	0.0289
1...	-0.0165	3	3	5	1	0.0724
2...	0	4	0	0	1	0
3...	-0.4746	5	2	0	0	1
4...	-0.2081	6	5	1	5	0.0675
5...	-0.4387	7	2	1	3	0.0631
7...	0	8	0	1	0	0
8...	0	9	0	1	0	0
9...	0	10	0	1	0	0
A...	-0.6578	11	2	0	1	0.1212
B...	0.8268	12	2	0	0	1
F...	0	13	0	1	0	0
G...	0	14	0	1	0	0
K...	0.3147	15	4	2	2	0.0492
L...	-0.0436	16	2	1	0	1
M...	1.3778	17	1	0	0	1
N...	0.5097	18	1	2	2	0.0364
P...	0.5975	19	1	0	1	0.0909
Q...	0	20	0	0	1	0
S...	-0.0732	21	1	2	5	0.0909
U...	0.0626	22	1	3	0	1
i...	0	23	0	1	0	0
u...	0	24	0	1	0	0
v...	0	25	0	1	0	0
w...	0.5219	26	2	5	3	0.047

(continued)

Table 9.6 (continued)

SA _k	CW(SA _k)	ID	N1	N2	N3	DEFECT[SA _k]
x...	0.3165	27	4	6	4	0.0195
y...	1.1738	28	2	0	0	1
z...	- 0.4104	29	3	1	4	0.0701

Threshold= 1 Number of SMILES Attributes (SA) = 29 Number of active SA = 19

Table 9.7 List of structural attributes (SA) and their correlation weights (CW) for the model (Eq. 9.6)

SA _k	CW(SA _k)	ID	N1	N2	N3	DEFECT[SA _k]
+...	1.2095	1	5	3	6	0.039
-...	1.0054	2	7	8	5	0.0273
0...	2.1456	3	1	0	0	1
1...	- 0.1480	4	4	4	0	1
2...	0.3166	5	1	0	0	1
3...	- 0.1899	6	2	0	1	0.1111
4...	- 0.4728	7	2	3	4	0.0438
5...	- 0.5935	8	1	1	5	0.1061
7...	0	9	0	1	0	0
8...	0	10	0	1	0	0
9...	0	11	0	0	1	0
A...	- 0.4671	12	2	1	0	1
B...	0.7469	13	2	1	0	1
F...	0	14	0	1	0	0
G...	0	15	0	1	0	0
K...	0.4414	16	4	2	2	0.0379
L...	0.3774	17	2	0	0	1
M...	1.1834	18	1	0	0	1
N...	0.6094	19	2	2	1	0.0364
P...	0.1610	20	1	0	1	0.0909
Q...	0	21	0	0	1	0
R...	0	22	0	0	1	0
S...	0.5237	23	1	1	5	0.1061
U...	- 0.1338	24	2	2	0	1
i...	0	25	0	1	0	0
u...	0	26	0	1	0	0
v...	1.5677	27	1	0	0	1
w...	- 0.3257	28	3	5	3	0.0372

(continued)

Table 9.7 (continued)

SA _k	CW(SA _k)	ID	N1	N2	N3	DEFECT[SA _k]
x...	0.0734	29	5	4	2	0.0427
y...	0.1315	30	3	0	0	1
z...	- 0.3574	31	1	2	6	0.1027

Threshold=1 Number of SMILES Attributes (SA) = 31 Number of active SA = 22

Table 9.8 List of structural attributes (SA) and their correlation weights (CW) for the model (Eq. 9.7)

SA _k	CW(SA _k)	ID	N1	N2	N3	DEFECT[SA _k]
+...	0.3146	1	5	3	4	0.0303
-...	0.2657	2	6	8	8	0.0165
0...	0	3	0	1	0	0
1...	- 0.0708	4	3	3	2	0.0265
2...	0	5	0	0	1	0
3...	- 0.4807	6	2	2	0	1
4...	- 0.3191	7	2	2	7	0.073
5...	- 0.2881	8	2	2	1	0.0394
7...	0	9	0	0	1	0
8...	1.6126	10	1	0	0	1
9...	- 0.7815	11	1	0	0	1
A...	- 0.5985	12	1	1	1	0.0051
B...	0.3118	13	1	2	0	1
F...	0	14	0	1	0	0
G...	1.1463	15	1	0	0	1
I...	0	16	0	0	1	0
K...	0.1807	17	5	1	3	0.0808
L...	0.1605	18	1	0	1	0.0909
M...	0	19	0	1	0	0
N...	0.1976	20	2	2	1	0.0394
P...	0.2354	21	1	1	1	0.0051
Q...	0	22	0	1	0	0
S...	- 0.098	23	1	2	4	0.0693
U...	0.9848	24	2	0	0	1
i...	0	25	0	1	0	0
w...	0.1925	26	2	5	5	0.0455
x...	0.0387	27	4	2	6	0.053
y...	0.8251	28	2	1	0	1
z...	- 0.1994	29	3	3	1	0.0541

Threshold=1 Number of SMILES Attributes (SA) = 29 Number of active SA = 21

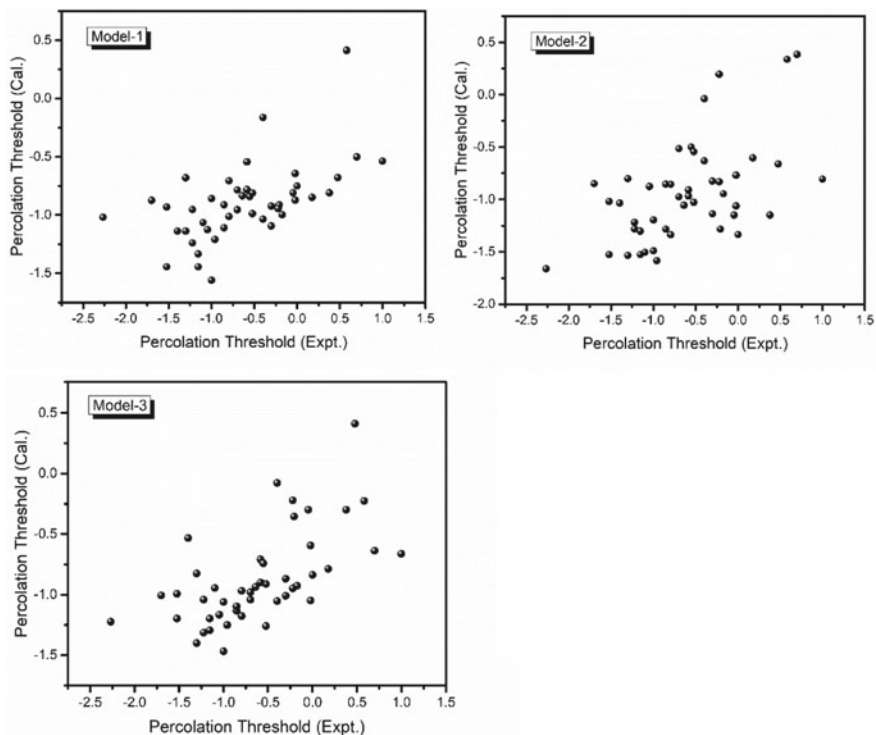


Fig. 9.2 Experimental and calculated percolation threshold valued for mathematical Model-1, Model-2, and Model-3

objects [93]. The development of the models for stochastic aspects of polymer agents' influence in medicine also is necessary for QSPR/QSAR-researches fields [94, 95].

9.8 Conclusion

The critical volume fraction that causes polymers to change from insulators to conductors is known as the 'percolation threshold'. The experimental percolation threshold cured data of 45 conductive polymer composite systems used in the present article were quite good and gave better performance when it was divided into four groups: active training set, passive training set, calibration set, and validation set. The suggested approach based on the quasi-SMILES, which are analogous to the traditional SMILES, gives reasonably good predictions for the percolation threshold for the studied conducting polymer composites (CPCs). The stability and reliability of the reported mathematical models are found to be reasonably stable, which is evident from the statistical parameters obtained for three random splits. The described

methodology is believed to be universal for similar situations where one aims to predict the response of an eclectic system upon a variety of physicochemical and/or biochemical conditions. The numerous conductive polymer variants, characteristics, conduction mechanisms, synthesis methods, and applications in diverse fields are also briefly highlighted in this communication.

References

1. Tijsanen J, Vlasveld D, Vuorinen J (2012) *Compos Sci Technol* 72(14):1741–1752. <https://doi.org/10.1016/j.compscitech.2012.07.009>
2. Mechrez G, Suckeveriene RY, Zelikman E, Rosen J, Ariel-Sternberg N, Cohen R, Narkis M, Segal E (2012) *ACS Macro Lett* 1(7):848–852. <https://doi.org/10.1021/mz300145a>
3. Al-Saleh MH, Sundararaj U (2009) *Carbon NY* 47(1):2–22. <https://doi.org/10.1016/j.carbon.2008.09.039>
4. Deng H, Lin L, Ji M, Zhang S, Yang M, Fu Q (2014) *Prog Polym Sci* 39(4):627–655. <https://doi.org/10.1016/j.progpolymsci.2013.07.007>
5. Stankovich S, Dikin DA, Dommett GHB, Kohlhaas KM, Zimney EJ, Stach EA, Piner RD, Nguyen SBT, Ruoff RS (2006) *Nature* 442(7100):282–286. <https://doi.org/10.1038/nature04969>
6. Ma P-C, Siddiqui NA, Marom G, Kim J-K (2010) *Compos Part A Appl Sci Manuf* 41(10):1345–1367. <https://doi.org/10.1016/j.compositesa.2010.07.003>
7. Villmow T, Pegel S, John A, Rentenberger R, Pötschke P (2011) *Mater Today* 14(7):340–345. [https://doi.org/10.1016/S1369-7021\(11\)70164-X](https://doi.org/10.1016/S1369-7021(11)70164-X)
8. Antunes RA, de Oliveira MCL, Ett G, Ett V (2011) *J Power Sources* 196(6):2945–2961. <https://doi.org/10.1016/j.jpowsour.2010.12.041>
9. Dang Z-M, Yuan J-K, Zha J-W, Zhou T, Li S-T, Hu G-H (2012) *Prog Mater Sci* 57(4):660–723. <https://doi.org/10.1016/j.pmatsci.2011.08.001>
10. Xu S, Rezvaniyan O, Peters K, Zikry MA (2013) *Nanotechnology* 24(15):155706. <https://doi.org/10.1088/0957-4484/24/15/155706>
11. Bauhofer W, Kovacs JZ (2009) *Compos Sci Technol* 69(10):1486–1498. <https://doi.org/10.1016/j.compscitech.2008.06.018>
12. Kirkpatrick S (1973) *Rev Mod Phys* 45(4):574. <https://doi.org/10.1103/RevModPhys.45.574>
13. Balberg I, Binenbaum N (1983) *Phys Rev B* 28(7):3799. <https://doi.org/10.1103/PhysRevB.28.3799>
14. Tang H, Chen X, Luo Y (1996) *Eur Polym J* 32(8):963–966. [https://doi.org/10.1016/0014-3057\(96\)00026-2](https://doi.org/10.1016/0014-3057(96)00026-2)
15. Scher H, Zallen R (1970) *J Chem Phys* 53(9):3759–3761. <https://doi.org/10.1063/1.1674565>
16. Grady BP (2010) *Macromol Rapid Commun* 31(3):247–257. <https://doi.org/10.1002/marc.200900514>
17. Gödel A, Kasaliwal G, Pötschke P (2009) *Macromol Rapid Commun* 30:423–429. <https://doi.org/10.1002/marc.200800549>
18. Grunlan JC, Gerberich WW, Francis LF (2001) *J Appl Polym Sci* 80:692–705. [https://doi.org/10.1002/1097-4628\(20010425\)80:4%3c692::AID-APP1146%3e3.0.CO;2-W](https://doi.org/10.1002/1097-4628(20010425)80:4%3c692::AID-APP1146%3e3.0.CO;2-W)
19. Pang H, Chen C, Bao Y, Chen J, Ji X, Lei J, Li Z-M (2012) *Mater Lett* 79:96–99. <https://doi.org/10.1016/j.matlet.2012.03.111>
20. Al-Saleh MH, Sundararaj U (2008) *Compos Part A Appl Sci Manuf* 39(2):284–293. <https://doi.org/10.1016/j.compositesa.2007.10.010>
21. Gupta S, Ou R, Gerhardt RA (2006) *J Electron Mater* 35(2):224–229. <https://doi.org/10.1007/BF02692439>
22. Malliaris A, Turner DT (1971) *J Appl Phys* 42(2):614–618. <https://doi.org/10.1063/1.1660071>

23. Kusy RP, Turner DT (1973) *J Appl Polym Sci* 17:1631–1633. <https://doi.org/10.1002/app.1973.070170528>
24. Du J, Zhao L, Zeng Y, Zhang L, Li F, Liu P, Liu C (2011) *Carbon NY* 49(4):1094–1100. <https://doi.org/10.1016/j.carbon.2010.11.013>
25. Zhang C, Ma C-A, Wang P, Sumita M (2005) *Carbon NY* 43(12):2544–2553. <https://doi.org/10.1016/j.carbon.2005.05.006>
26. Al-Saleh MH, Jawad SA, El Ghanem HM (2014) *High Perform Polym* 26(2):205–211. <https://doi.org/10.1177/0954008313507590>
27. Wang B, Li H, Li L, Chen P, Wang Z, Gu Q (2013) *Compos Sci Technol* 89:180–185. <https://doi.org/10.1016/j.compscitech.2013.10.002>
28. Grunlan JC, Gerberich WW, Francis LF (2001) *Polym Eng Sci* 41:1947–1962. <https://doi.org/10.1002/pen.10891>
29. Jurewicz I, King AAK, Worajittiphon P, Asanithi P, Brunner EW, Sear RP, Hosea TJC, Keddie JL, Dalton AB (2010) *Macromol Rapid Commun* 31:609–615. <https://doi.org/10.1002/marc.200900799>
30. Grossiord N, Loos J, Koning CE (2005) *J Mater Chem* 15(24):2349–2352. <https://doi.org/10.1039/B501805F>
31. Kyrylyuk A, Hermant M, Schilling T, Klumperman B, Koning CE, der Schoot P (2011) *Nature Nanotech* 6:364–369. <https://doi.org/10.1038/nnano.2011.40>
32. Chen J, Shi Y-Y, Yang J-H, Zhang N, Huang T, Chen C, Wang Y, Zhou Z-W (2012) *J Mater Chem* 22(42):22398–22404. <https://doi.org/10.1039/C2JM34295B>
33. Gubbels F, Jérôme R, Vanlathem E, Deltour R, Blacher S, Brouers F (1998) *Chem Mater* 10(5):1227–1235. <https://doi.org/10.1021/cm970594d>
34. Dai K, Xu X-B, Li Z-M (2007) *Polymer* 48(3):849–859. <https://doi.org/10.1016/j.polymer.2006.12.026>
35. Zhang Y-C, Dai K, Tang J-H, Ji X, Li Z-M (2010) *Mater Lett* 64(13):1430–1432. <https://doi.org/10.1016/j.matlet.2010.03.041>
36. Breuer O, Tchoudakov R, Narkis M, Siegmann A (2000) *Polym Eng Sci* 40:1015–1024. <https://doi.org/10.1002/pen.11229>
37. Jurewicz I, Worajittiphon P, King AAK, Sellin PJ, Keddie JL, Dalton AB (2011) *J Phys Chem B* 115(20):6395–6400. <https://doi.org/10.1021/jp111998p>
38. Regev O, ElKati P, Loos J, Koning C (2004) *Adv Mater* 16:248–251. <https://doi.org/10.1002/adma.200305728>
39. Linares A, Canalda JC, Cagiao ME, Ezquerria TA (2011) *Compos Sci Technol* 71(10):1348–1352. <https://doi.org/10.1016/j.compscitech.2011.05.008>
40. Pham VH, Dang TT, Hur SH, Kim EJ, Chung JS (2012) *ACS Appl Mater Interfaces* 4(5):2630–2636. <https://doi.org/10.1021/am300297j>
41. Yu J, Lu K, Sourty E, Grossiord N, Koning CE, Loos J (2007) *Carbon NY* 45(15):2897–2903. <https://doi.org/10.1016/j.carbon.2007.10.005>
42. Zhan Y, Lavorgna M, Buonocore G, Xia H (2012) *J Mater Chem* 22(21):10464–10468. <https://doi.org/10.1039/C2JM31293J>
43. Tkalya EE, Ghislandi M, de With G, Koning CE (2012) *Curr Opin Colloid Interface Sci* 17(4):225–232. <https://doi.org/10.1016/j.cocis.2012.03.001>
44. Ghislandi M, Tkalya E, Marinho B, Koning CE, de With G (2013) *Compos Part A Appl Sci Manuf* 53:145–151. <https://doi.org/10.1016/j.compositesa.2013.06.008>
45. Hermant MC, Klumperman B, Kyrylyuk AV, van der Schoot P, Koning CE (2009) *Soft Matter* 5(4):878–885. <https://doi.org/10.1039/B814976C>
46. Yu C, Choi K, Yin L, Grunlan JC (2011) *ACS Nano* 5(10):7885–7892. <https://doi.org/10.1021/nn202868a>
47. Hermant M-C, van der Schoot P, Klumperman B, Koning CE (2010) *ACS Nano* 4(4):2242–2248. <https://doi.org/10.1021/nn901643h>
48. Balberg I, Azulay D, Toker D, Millo O (2004) *Int J Mod Phys B* 18(15):2091–2121. <https://doi.org/10.1142/S0217979204025336>

49. Gao J-F, Li Z-M, Meng Q-J, Yang Q (2008) *Mater Lett* 62(20):3530–3532. <https://doi.org/10.1016/j.matlet.2008.03.053>
50. Celzard A, McRae E, Deleuze C, Dufort M, Furdin G, Marêché JF (1996) *Phys Rev B* 53(10):6209. <https://doi.org/10.1103/PhysRevB.53.6209>
51. Grossiord N, Loos J, van Laake L, Maugey M, Zakri C, Koning CE, Hart AJ (2008) *Adv Funct Mater* 18:3226–3234. <https://doi.org/10.1002/adfm.200800528>
52. Grunlan J, Mehrahi A, Bannon M, Bahr J (2004) *Adv Mater* 16:150–153. <https://doi.org/10.1002/adma.200305409>
53. Mierczynska A, Mayne-L'Hermite M, Boiteux G, Jeszka JK (2007) *J Appl Polym Sci* 105:158–168. <https://doi.org/10.1002/app.26044>
54. Tan Y, Fang L, Xiao J, Song Y, Zheng Q (2013) *Polym Chem* 4(10):2939–2944. <https://doi.org/10.1039/C3PY00164D>
55. Al-Saleh MH, Gelves GA, Sundararaj U (2011) *Compos Part A Appl Sci Manuf* 42(1):92–97. <https://doi.org/10.1016/j.compositesa.2010.10.003>
56. Mamunya YP, Davydenko VV, Pissis P, Lebedev EV (2002) *Eur Polym J* 38(9):1887–1897. [https://doi.org/10.1016/S0014-3057\(02\)00064-2](https://doi.org/10.1016/S0014-3057(02)00064-2)
57. Gelves GA, Al-Saleh MH, Sundararaj U (2011) *J Mater Chem* 21(3):829–836. <https://doi.org/10.1039/C0JM02546A>
58. Sharma PS, Pietrzyk-Le A, D'Souza F, Kutner W (2012) *Anal Bioanal Chem* 402(10):3177–3204. <https://doi.org/10.1007/s00216-011-5696-6>
59. Evans P, Grigg R, Monteith M (1999) *Tetrahedron Lett* 40(28):5247–5250. [https://doi.org/10.1016/S0040-4039\(99\)00993-4](https://doi.org/10.1016/S0040-4039(99)00993-4)
60. Masuda T, Karim SMA, Nomura R (2000) *J Mol Catal A Chem* 160(1):125–131. [https://doi.org/10.1016/S1381-1169\(00\)00239-9](https://doi.org/10.1016/S1381-1169(00)00239-9)
61. Deronzier A, Moutet J-C (1996) *Coord Chem Rev* 147:339–371. [https://doi.org/10.1016/0010-8545\(95\)01130-7](https://doi.org/10.1016/0010-8545(95)01130-7)
62. Guay J, Diaz A, Wu R, Tour JM, Dao LH (1992) *Chem Mater* 4(2):254–255. <https://doi.org/10.1021/cm00020a006>
63. Gomes AL, Zakia MBP, Godoy Filho J, Armelin E, Aleman C, de Carvalho Campos JS (2012) *Polym Chem* 3(5):1334–1343. <https://doi.org/10.1039/C2PY00003B>
64. Arefi F, Andre V, Montazer-Rahmati P, Amouroux J (1992) *Pure Appl Chem* 64(5):715–723. <https://doi.org/10.1351/pac199264050715>
65. Staiti P, Lufrano F (2005) *J Electrochem Soc* 152(3):A617. <https://doi.org/10.1149/1.1859614>
66. Miyata M, Noma F, Okanishi K, Tsutsumi H, Takemoto K (1987) In: Atwood JL, Davies JED (eds) *Inclusion phenomena in inorganic, organic, and organometallic hosts*. *Adv Incl Sci* 4:249–252. https://doi.org/10.1007/978-94-009-3987-5_42
67. Long Y, Chen Z, Shen J, Zhang Z, Zhang L, Xiao H, Wan M, Duvail JL (2006) *J Phys Chem B* 110(46):23228–23233. <https://doi.org/10.1021/jp062262e>
68. Bajpai M, Srivastava R, Dhar R, Tiwari RS (2016) *Indian J Mater Sci* 2016:5842763. <https://doi.org/10.1155/2016/5842763>
69. Miasik JJ, Hooper A, Tofield BC (1986) *J Chem Soc Faraday Trans 1 Phys Chem Condens Phases* 82(4):1117–1126. <https://doi.org/10.1039/F1986820117>
70. Malhotra BD, Singhal R, Chaubey A, Sharma SK, Kumar A (2005) *Curr Appl Phys* 5(2):92–97. <https://doi.org/10.1016/j.cap.2004.06.021>
71. Umana M, Waller J (1986) *Anal Chem* 58(14):2979–2983. <https://doi.org/10.1021/ac00127a018>
72. Zhu Z, Wang G, Sun M, Li X, Li C (2011) *Electrochim Acta* 56(3):1366–1372. <https://doi.org/10.1016/j.electacta.2010.10.070>
73. Sultana I, Rahman MM, Wang J, Wang C, Wallace GG, Liu H-K (2012) *Electrochim Acta* 83:209–215. <https://doi.org/10.1016/j.electacta.2012.08.043>
74. Pang H, Xu L, Yan D-X, Li Z-M (2014) *Prog Polym Sci* 39(11):1908–1933. <https://doi.org/10.1016/j.progpolymsci.2014.07.007>
75. Toropov AA, Toropova AP (2015) *Chemosphere* 124:40–46. <https://doi.org/10.1016/j.chemosphere.2014.10.067>

76. Toropov AA, Toropova AP (2015) *Chemosphere* 139:18–22. <https://doi.org/10.1016/j.chemosphere.2015.05.042>
77. Katritzky AR, Petrukhin R, Jain R, Karelson M (2001) *J Chem Inf Comput Sci* 41(6):1521–1530. <https://doi.org/10.1021/ci010043e>
78. Khajeh A, Rasaei MR (2012) *Struct Chem* 23(2):399–406. <https://doi.org/10.1007/s11224-011-9879-8>
79. Afantitis A, Melagraki G, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2006) *Polymer* 47(9):3240–3248. <https://doi.org/10.1016/j.polymer.2006.02.060>
80. Toropova AP, Toropov AA, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2011) *J Comput Chem* 32(12):2727–2733. <https://doi.org/10.1002/jcc.21848>
81. Toropov AA, Toropova AP, Begum S, Achary PGR (2016) *SAR QSAR Environ Res* 27(4):293–301. <https://doi.org/10.1080/1062936X.2016.1172666>
82. Golzar K, Amjad-Iranagh S, Modarress H (2013) *Measurement* 46(10):4206–4225. <https://doi.org/10.1016/j.measurement.2013.08.012>
83. Taniwaki H, Kaneko H (2022) *Polym Eng Sci* 62(9):2750–2756. <https://doi.org/10.1002/pen.26058>
84. Huang R, Siddiqui MK, Manzoor S, Khalid S, Almotairi S (2022) *Eur Phys J Plus* 137(3):410. <https://doi.org/10.1140/epjp/s13360-022-02629-3>
85. Ishikiriyama K (2022) *Thermochim Acta* 708:179135. <https://doi.org/10.1016/j.tca.2021.179135>
86. Khan PM, Roy K (2022) *Mol Inform* 41(1):2000030. <https://doi.org/10.1002/minf.202000030>
87. Schustik SA, Cravero F, Ponzoni I, Díaz MF (2021) *Comput Mater Sci* 194:110460. <https://doi.org/10.1016/j.commatsci.2021.110460>
88. Karuth A, Alesadi A, Xia W, Rasulev B (2021) *Polymer* 218:123495. <https://doi.org/10.1016/j.polymer.2021.123495>
89. Wang S, Cheng M, Zhou L, Dai Y, Dang Y, Ji X (2021) *SAR QSAR Environ Res* 32(5):379–393. <https://doi.org/10.1080/1062936X.2021.1902387>
90. Shao S, Sun H, Muhammad Y, Huang H, Wang R, Nie S, Huang M, Zhao Z, Zhao Z (2021) *Food Res Int* 141:110144. <https://doi.org/10.1016/j.foodres.2021.110144>
91. Gupta S, Mallick S (2018) *SAR QSAR Environ Res* 29(3):171–186. <https://doi.org/10.1080/1062936X.2017.1419985>
92. Mallakpour S, Hatami M, Golmohammadi H (2011) *J Mol Model* 17(7):1743–1753. <https://doi.org/10.1007/s00894-010-0885-3>
93. Holder AJ, Ye L, Eick JD, Chappelow CC (2006) *QSAR Comb Sci* 25(10):905–911. <https://doi.org/10.1002/qsar.200510203>
94. González-Díaz H, Pérez-Bello A, Uriarte E (2005) *Polymer* 46(17):6461–6473. <https://doi.org/10.1016/j.polymer.2005.04.104>
95. González-Díaz H, Saiz-Urra L, Molina R, Uriarte E (2005) *Polymer* 46(8):2791–2798. <https://doi.org/10.1016/j.polymer.2005.01.066>

Chapter 10

On the Possibility to Build up the QSAR Model of Different Kinds of Inhibitory Activity for a Large List of Human Intestinal Transporter Using Quasi-SMILES



P. Ganga Raju Achary, P. Kali Krishna, Alla P. Toropova,
and Andrey A. Toropov

Abstract Membrane transporters play a significant role in pharmacokinetics and drug resistance and mediate many biological effects of substances. Among biologically active chemicals, it is necessary to evaluate the profiles of their transporter interactions in order to identify potential medication candidates. The constraints and predictive capability of models for substances with heterogeneous physicochemistry and variable permeability/absorption are explored in this communication using the largest diverse permeability and absorption dataset for 3199 compounds. Here, we offer a classification-based QSAR model of different inhibitory activities for an extensive list of Human Intestinal Transporter using quasi-SMILES. The extraction of properties from quasi-SMILES and the computation of so-called correlation weights for these attributes using Monte Carlo techniques were the foundation for the classification-based models. As qualitative statistical validation criteria, the classification model was tested using sensitivity (= 0.86), specificity (= 1), accuracy (= 0.96), and Matthews correlation coefficient (MCC = 0.90). Described computational experiments confirm the suitability of application of so-called Index of Ideality of Correlation to improve the predictive potential of the models.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-28401-4_10.

P. Ganga Raju Achary (✉)
Department of Chemistry, Institute of Technical Education and Research (ITER), Siksha 'O' Anusandhan University, Bhubaneswar, Odisha 751030, India
e-mail: pgrachary@soa.ac.in

P. Kali Krishna
Department of Bioinformatics, B.J.B Autonomous College, Bhubaneswar, Odisha, India

A. P. Toropova · A. A. Toropov
Laboratory of Environmental Chemistry and Toxicology, Department of Environmental Health Science, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156 Milano, Italy

Keywords Human intestinal transporter · Classification model · Quasi-SMILES · Index of ideality of correlation

10.1 Introduction

Even though oral medication delivery is the preferred method at the moment, intestinal drug absorption is hampered by a number of highly variable and unpredictable processes, including gastrointestinal motility, intestinal drug solubility, and intestinal metabolism. The intestinal drug transport, which is mediated by many transmembrane proteins, including P-glicoprotein (P-gp), breast cancer resistance protein (BCRP), human peptide transporter 1 (PEPT1), and organic anion transporting polypeptide 2B1 (OATP2B1), is another factor that has been discovered and characterised over the past 20 years. It is generally known that intestinal transporters have a substantial impact on the oral absorption of many medications, either by promoting their cellular uptake or by pumping the medications back to the gut lumen, which reduces the oral bioavailability of the pharmaceuticals. When medications that elicit transporter induction or inhibition are given concurrently with other pharmaceuticals, the functional relevance of these drugs becomes even more clear in cases of unintended drug-drug interactions, which, in turn, affects the number of drugs exposed. The preferred site of intestinal medication absorption may be affected functionally by the non-homogeneous longitudinal expression of a number of intestinal transporters along the human intestine. Understanding the precise location of pharmacologically important transporters on the apical or basolateral membrane of enterocytes, which is occasionally disputed, is also of importance. Furthermore, there is clearly a connection between intestine transporters (apical-basolateral), intestinal enzymes and transporters, and intestinal and hepatic transporters.

For the development of new drugs, intestinal absorption prediction models are essential. For more common “drug-like” compounds (also known as “rule of 5” or Ro5 drugs), there are numerous *in silico*, *in vitro*, and *in vivo* models available to help and understand the process in a better way. However, there are many concerns regarding the applicability of these models to “non-drug-like” compounds typically found for “undruggable targets” (also known as “beyond the rule of 5” or “bRo5 drugs”) [1–6]. Given that these drugs are frequently bigger, more complicated, and have lower permeability, there are concerns about the applicability of such models in this area. Medicinal researchers are uncomfortable using current or existing models that haven’t been thoroughly tested for intestinal absorption in these circumstances [6].

All biological species have membrane transport proteins, such as members of the ATP-binding cassette (ABC) superfamily and the solute carrier (SLC) family. By regulating their cellular inflow or efflux, transporters are known to impact the membrane permeability of numerous xenobiotic and endogenous substances [7]. Active transport proteins (like P-glycoprotein) are crucial for pharmacokinetics, drug-drug interactions, and multidrug resistance [8, 9]. Membrane transporters are

widely distributed and expressed across all tissues and organs; the human genome only contains about 600 transport proteins from the ABC and SLC families [10, 11]. When one takes into account the interdependencies between membrane transporters as well as their relationships with other metabolic systems, one can see the actual intricacy of processes governing the permeability of tiny molecules across biological membranes [7, 12].

Absorption, distribution, metabolism, and elimination influence how drugs are disposed of (ADME). Understanding absorption is important because it affects total systemic exposure, and oral administration is the most frequent way to provide small compounds. The ability of a substance to pass through the intestinal wall (f_a) and avoid intestinal and hepatic metabolism (F_g and F_h , respectively) determines how much of it can be absorbed orally. Total oral bioavailability (F), the term used to describe the total fraction of a drug that reaches the systemic circulation, is a function of three factors (Eq. 10.1).

$$F = f_a * F_g * F_h \quad (10.1)$$

The susceptibility of a molecule to first-pass metabolism may have an impact on the compound's overall oral bioavailability. When the bioavailability in the stomach and liver is known, f_a can be calculated and used as a “cleaner” metric to assess human absorption.

$$f_a = \frac{F}{F_g * F_h} \quad (10.2)$$

Computer simulations have been used for in silico quantitative structure–activity relationship (QSAR) [12] models to realise how a compound's chemical structure affects its ADME qualities. Over time, numerous QSAR techniques have been created to learn which chemical characteristics can enhance absorption through a variety of endpoints, including Caco-2 permeability, effective human permeability [12, 13], and first-order human rate of absorption (K_a) [14]. Before even being produced in the lab, these correlations can assist the design of novel molecules with enhanced absorption properties.

The present book chapter aims to offer a distinctive viewpoint on the usefulness of the most recent and well-liked models for forecasting human f_a for *bRo5* and low permeability/absorption organic molecules. We can assess the constraints and predictive capability of models for substances with heterogeneous physicochemistry and variable permeability/absorption using the largest permeability and absorption dataset compiled to date (to our knowledge, with $n = 3199$). Here, we offer a classification-based QSAR model of different kinds of inhibitory activity for a large list of Human Intestinal Transporter using quasi-SMILES [15–20]. The suggested models were obtained by the Monte Carlo technique via the CORAL software (<http://www.insilico.eu/coral>) with applying the quasi-SMILES technology.

10.1.1 Literature Review on Various QSAR Models for Human Intestinal Transporter

A hierarchical support vector regression-based *in silico* model for Caco-2 permeability is being reported. One of the important considerations in the process of discovering and developing new drugs is drug absorption. To begin studying intestinal absorption, the human colon cancer cell layer (Caco-2) model has commonly been used as a surrogate. A novel machine learning-based hierarchical support vector regression (HSVR) method was used to create a QSAR model to represent the highly confusing passive diffusion and transporter-mediated active transport. The experimental values of the training samples, test samples, and outlier samples showed a high degree of agreement with the HSVR model. A mock test and a number of rigorous statistical standards were used to validate further and confirm the predictability of HSVR. In order to aid in the creation of new drugs, this HSVR model can be used to predict the Caco-2 permeability [21].

A vast class of polyphenols known as flavonoids is present in a wide variety of plant-based meals. While flavonoids have a variety of biological properties, including anti-cancer, antioxidant, and anti-inflammatory properties, their poor oral bioavailability has been viewed as a significant barrier to their utilisation as functional foods. The bioavailability of flavonoids is affected by cellular absorption and efflux.

Twenty-seven flavonoids were assessed for their cellular absorption in Caco-2 cells with verapamil and cellular uptake of flavonoids without verapamil to research their cellular uptake and efflux. Then, from each compound's matching without verapamil, a quantitative structure-absorption relationship (QSAR) model was constructed. The model had a high cross-validation coefficient (Q^2) value of 0.809 and showed good resilience and predictability [22].

Flavonoid interactions during digestion, absorption, distribution, and metabolism: a sequential QSAR-based approach has been carried out in the study of bioavailability and bioactivity. When consumed, the group of polyphenols known as flavonoids promotes good health. However, their low bioavailability is a significant barrier to their usage as medications or nutraceuticals. Flavonoid interactions at digestion, absorption, and distribution phases have been linked to low bioavailability, and their molecular structure significantly impacts these interactions [23].

Critical evaluation of human oral bioavailability for pharmaceutical drugs is carried out by using various cheminformatics approaches. In clinical trials, a novel drug's oral bioavailability (%F) is a critical element that influences its outcome. Historically, expensive, and time-consuming experimental tests have been used to determine %F. In order to improve the drug development process, computational models that assess potential drugs' %F properties before they are manufactured should be created. To create a number of computational %F models, researchers used a combinatorial QSAR technique. A dataset of 995 medications is from open sources. Chemical descriptors for each drug were created, and the appropriate QSAR models were created using random forest, support vector machine, k closest neighbour, and CASE Ultra. Fivefold cross-validation was used to validate the models that

were generated. The reliability of %F values' external predictivity was low ($R^2 = 0.28$, $n = 995$, $MAE = 24$), but it was enhanced ($R^2 = 0.40$, $n = 362$, $MAE = 21$) by removing unreliable predictions that were highly likely to interact with MDR1 and MRP2 transporters. A further outcome of categorising the compounds using the %F values (%F 50% as "low", %F 50% as "high") and creating category QSAR models was an external accuracy of 76%. The integration of data on drug-transporter interactions considerably improves the predictive %F QSAR models that were constructed, which might be utilised to assess new therapeutic compounds [24].

Structural determinants for transport across the intestinal bile acid transporter using C-24 bile acid conjugates are also reported. The human apical sodium-dependent bile acid transporter (hASBT) is a potential prodrug target to improve oral drug absorption and reabsorption of bile acid per day. Cross-validation was used to assess the CSP-SAR models, which were developed using structural and physicochemical descriptors. One structural and three physicochemical descriptors were used in the best CSP-SAR model for $K_m/normV_{max}$, which similarly showed that hydrophobicity improved efficiency [25].

Computational models for drug inhibition of the human apical sodium-dependent bile acid transporter are carried out. The human apical sodium-dependent bile acid transporter (ASBT; SLC10A2) is the main mechanism for intestinal bile acid reabsorption. Secondary bile acids raise the danger of colon cancer. As a result, medications that block ASBT may raise the risk of colon cancer. The authors aimed through this work to develop computational models for ASBT inhibition and to discover FDA-approved medications that inhibit ASBT [26, 27]. A modified Laplacian Bayesian modelling method using 2D descriptors, a HipHop qualitative approach, and a Hypogen quantitative approach were all used in computational modelling. Thirty substances were first tested for ASBT inhibition. The most powerful 11 molecules were used to create a qualitative pharmacophore, which was then used to search a drug database, producing 58 hits. The K_i values of other substances were evaluated after testing. Using 38 compounds, a 3D-QSAR and a Bayesian model were created. According to a validation examination, both models have shown good predictability in determining whether a medicine is a powerful or non-potent ASBT inhibitor. The most effective chemicals were appropriately rated by the Bayesian model. It was discovered that many FDA-approved medications from various families, including dihydropyridine calcium channel blockers and HMG CoA-reductase inhibitors, are ASBT inhibitors utilising a combined in vitro and computational method [26, 27].

A QSAR study for the translocation of tripeptides via the human proton-coupled peptide transporter, hPEPT1 (SLC15A1), is reported. It has been discovered that the human intestine proton-coupled peptide transporter, hPEPT1 (SLC15A1), functions as an absorptive transporter for both prodrugs and drug molecules. Models based on competitive tests have so far helped to grasp the conditions for transport. The predictive power of these models for substrate translocation via hPEPT1 is rather low. The study's objective was to look into the prerequisites for translocation via hPEPT1. Using a statistical approach, a set of 55 tripeptides was chosen using a principal component analysis based on VolSurf descriptors. A large portion of these tripeptides has not yet been studied. An MDCK/hPEPT1 cell-based translocation assay assessing

substrate-induced variations in the fluorescence of a membrane potential-sensitive probe was used to quantify the tripeptides' translocation via hPEPT1. Competition experiments with [¹⁴C]Gly-Sar in MDCK/hPEPT1 cells were used to evaluate the affinities of pertinent tripeptides for hPEPT1. It was discovered that forty tripeptides were hPEPT1 substrates, with K_m app values ranging from 0.4 to 28 mM. A QSAR model connecting K_m app values with VolSurf descriptors was built to rationalise the need for transportation. This is the first prediction model for the hPEPT1-mediated translocation of tripeptides [28].

The discovery of ligands for the human intestinal di-/tripeptide transporter (hPEPT1) was carried out using a QSAR-assisted virtual screening strategy [29–31].

SAR models were proposed for the binding of tripeptides and tripeptidomimetics to the human intestinal di-/tripeptide transporter hPEPT1. 3D-QSAR models were built based on a series of 25 different tripeptides for the binding of tripeptides and tripeptidomimetics to hPEPT1. By using multivariate data analysis, VolSurf descriptors were created and associated with binding affinities. Using Caco-2 cell monolayers, tripeptides and tripeptidomimetics have their affinities for hPEPT1 experimentally evaluated. The structural variety of the 25 tripeptides and tripeptidomimetics was defined by VolSurf descriptors, and their K_i values ranged from 0.15 to 25 mM. A QSAR model was created to connect the tripeptides' experimentally determined binding affinity for hPEPT1 with their VolSurf characteristics. The QSAR model was used to derive structural data on tripeptide characteristics impacting the binding to hPEPT1. This knowledge could be useful for developing tripeptides and tripeptidomimetics that target hPEPT1 as an absorptive transporter to enhance intestinal absorption [30].

The dipeptide model suggested the intestinal oligopeptide transporter. By creating peptidomimetic prodrugs, it has been proposed that the human intestinal di-/tripeptide carrier, hPepT1, could be a drug delivery target for enhancing intestinal transport of poor permeability substances. These findings suggest that the dipeptide prodrug principle is a promising drug delivery paradigm. It has been demonstrated that model ester prodrugs use D-Glu-Ala and D-Asp-Ala as pro-moieties for benzyl alcohol maintain an affinity for hPepT1. D-Asp(BnO)-Ala and D-trans epithelial Glu(BnO)-Ala's transport investigations in Caco-2 cells revealed that the K_m for trans epithelial transport was not significantly different for the two compounds. Additionally, there is no difference in the maximum transport rate of the carrier-mediated flux component between the two model prodrugs [31].

The progress in predicting human ADME parameters by various *in silico* methods is being continuously given attention from time to time. Analysing the evolution of a scientific approach is a useful exercise for predicting the future course that the process might follow. There are distinct eras in the recent history of computational techniques to study absorption, distribution, metabolism, and excretion (ADME). With the work of Corwin Hansch and others, the first started in the 1960s and continued into the 1970s [32]. Small collections of *in vivo* ADME data were used in their models. The second period, which spanned the 1980s and 1990s, saw extensive use of *in vitro* methods as substitutes for *in vivo* ADME research. These strategies encouraged the development and expansion of interpretable computational ADME models that are

now widely available in the literature. The third era is now, and there are numerous literature datasets for absorption, drug-drug interactions (DDI), drug transporters and efflux pumps [P-gp, multidrug resistance protein (MRP)], intrinsic clearance, and brain penetration that are derived from in vitro data and can theoretically be used to predict the situation in vivo in humans.

Pharmaceutical corporations have been under constant pressure to accelerate drug discovery while lowering drug development costs, which has led to the emergence of combinatorial synthesis, high throughput screening, and computational techniques. Reduced drug candidate dropout rates are desired in drug development's final, most expensive phases. This is done by speeding up the nomination of likely clinical candidates and raising the failure rate of candidate molecules during the preclinical stages. The market is now aware that toxicity and pharmacokinetics are the primary causes of clinical failure aside from efficacy. In order to evaluate features such as metabolic stability, cytochrome P-450 inhibition, absorption, and genotoxicity earlier in the drug discovery paradigm, major firm investment in ADME and drug safety departments occurred in the late 1990s. Evaluating higher throughput data to see if computational (in silico) models can be built and verified from it is the natural next step in this process. With such models, the number of chemicals that could be virtually screened for ADME characteristics could expand exponentially. To address intestinal permeability and cytochrome P-450-mediated DDI, many researchers have begun to use in silico, in vitro, and in vivo techniques concurrently [33].

Another study uses comparative molecular field analysis (CoMFA), a three-dimensional method for developing QSAR, to examine the relationship between chemical structure (steric and electrostatic fields) and affinity for the small intestinal oligopeptide carrier (PepT1). Numerous chemical descriptors (CoMFA fields, isobutyl alcohol/water distribution coefficients, K_t , J_{max} , and P_c) and biological activity parameters (K_t , J_{max} , and P_c) were investigated. The regression line between the experimental and calculated P_c had a slope of 0.994 and an intercept of 0.009. The model suited the experimental data with a correlation coefficient of 0.993 and a standard error of 0.041. These findings improve our knowledge of the molecular prerequisites for ideal drug-carrier interactions with the intestinal peptide transporter and provide a helpful visual tool for developing novel, potentially intriguing structures that have an affinity for PepT1 [34].

In a comparative molecular field analysis, data from a number of bile acid analogues were used to create a link between structure and binding activity for the intestinal bile acid transporter (CoMFA). The investigated compounds included a number of bile acid-peptide conjugates with modifications at the cholic acid sterol nucleus position 24, as well as compounds with minor modifications at positions 3, 7, and 12. These substances were split into a training set and a test set for the CoMFA investigation, each consisting of 25 and 5 molecules, respectively. With a cross-validated, conventional, and predictive R^2 of 0.63, 0.96, and 0.69, respectively, the best three-dimensional QSAR model discovered rationalises the steric and electrostatic factors that modulate affinity to the bile acid carrier, indicating a good predictive model for carrier affinity. Positioning an electronegative moiety at the specified positions and adding steric bulk to the side chain's terminus help bind. The

model recommends replacements that could result in novel substrates with a suitable affinity for the carrier at the positions selected positions [35].

10.1.2 An Overview of Computer Simulations Study of Human Intestinal Transporter

An interesting in vivo, in situ, in vitro, and in silico studies report the influence of rhein on the absorption of rehmanioside D. Authors stated that breast cancer resistance and multidrug resistance-associated protein 2 affected the intestinal epithelium's permeability by mediating the stimulation of absorption of rehmanioside D in the presence or absence of rhein [36].

Physiologically based pharmacokinetic (PBPK) modelling presented to evaluate in vitro-to-in vivo extrapolation for intestinal P-glycoprotein (P-gp) inhibition. In order to quantitatively anticipate drug-drug interactions (DDIs) on drug-metabolising enzymes and transporters, PBPK modelling coupled with in vitro-to-in vivo extrapolation (IVIVE) is commonly used in model-informed drug discovery and development. Through the use of PBPK modelling, this study sought to examine an IVIVE for intestinal P-gp-mediated DDIs, including three P-gp substrates-digoxin, dabigatran etexilate, and quinidine- and two P-gp inhibitors-itraconazole and verapamil [37].

A comparative study on the intestinal absorption of three gastrodin analogues via the glucose transport pathway is reported in the paper [38]. Three gastrodin analogues, salicin, arbutin, and 4-methoxyphenyl-D-glucoside, have their intestinal absorption characteristics assessed using conventional biopharmaceutical and computer-aided molecular docking techniques (4-MG). The logP values of the gastrodin analogues were found to be in the following order: 4-MG > salicin > arbutin, according to the oil-water partition coefficient (logP) studies. Arbutin's apparent permeability coefficient value was found to be higher than that of salicin and 4-MG for in vitro Caco-2 cell transport studies. Arbutin and 4-MG were more effectively absorbed than salicin, according to in situ single-pass intestinal perfusion tests, and the three compounds were more effectively absorbed in the small intestine than the colon. Therefore, the difference in chemical structure can have an impact on absorption [38].

An in silico, in vitro, and ex vivo approach was presented for the intestinal efflux transporter inhibition activity of xanthenes from mangosteen pericarp [39].

PBPK model-informed drug development for fenebrutinib is presented to understand complex drug-drug interactions. In vitro, fenebrutinib inhibits BCRP and OATP1B transporters as well as CYP3A substrate and time-dependently. The ultimate goal of developing PBPK modelling methodologies was to comprehend complex drug-drug interactions (DDIs) and suggest doses for hypothetical situations. Because fenebrutinib inhibits intestine BCRP rather than hepatic OATP1B, the results of two separate methods: PBPK simulation and endogenous biomarker

measurement were consistent and supported this theory. The unexpected observation of itraconazole-fenebrutinib DDI (maximum plasma concentration (C_{max}) lowered, and area under the curve (AUC) increased) was explained by a mechanistic-absorption model that took into consideration the effects of excipient complexation with fenebrutinib. Overall data from clinical and nonclinical studies, sensitivity analyses, and other sources indicated that fenebrutinib is probably a sensitive CYP3A substrate. Without the need for additional clinical DDI trials, this enhanced PBPK application enabled the adoption of a model-informed approach to assist in the establishment of concomitant medicine recommendations for fenebrutinib [40].

Development of simplified *in vitro* P-Glycoprotein substrate assay and *in silico* prediction models was presented to evaluate the transport potential of P-gp. Simplifying P-gp substrate tests and offering *in silico* models that forecast P-transport gp's potential are essential for effective drug discovery and screening. The study aimed at creating a more straightforward *in vitro* screening approach to assess P-gp substrates in cells overexpressing P-gp via unidirectional membrane transport. Additionally, the test set's low-potential classes in the random forest three-class classification model displayed high balanced accuracy of 0.821 and precision of 0.761. Authors concluded that the streamlined *in vitro* P-gp substrate assay was appropriate for screening compounds in the early stages of drug discovery and that the *in silico* regression model and three-class classification model using only chemical structure information could identify the transport potential of compounds, including P-gp-mediated flux ratios. The approach is anticipated to be a useful tool to enhance efficient central nervous system medications and enhance intestine absorption [41].

Prebiotics and probiotics, which are combined to form synbiotics, may be utilised to treat diseases like colorectal cancer (CRC) by altering the human gut microbiota. The potential combinatorial mechanisms of action of such regimens have not yet been identified due to methodological restrictions. In order to co-culture CRC-derived epithelial cells with a model probiotic under a simulated prebiotic regimen, HuMiX gut-on-a-chip model was enlarged. Researchers also linked the multi-omic data with *in silico* metabolic modelling. In contrast to separate prebiotic or probiotic treatments, the synbiotic regimen decreased levels of the oncometabolite lactate and downregulated genes involved in drug resistance and procarcinogenic pathways. The simulated regimens resulted in various ratios of organic and short-chain fatty acids being generated. The synbiotic diet was applied to primary CRC-derived cells, which resulted in a diminished capacity for self-renewal. This strategy exemplifies the promise of modelling for logically developing medicines based on synbiotics in future [42].

Computational discovery and experimental validation of inhibitors of the Human Intestinal Transporter OATP2B1 are elaborated on in the article [43]. Human organic anion transporters (OATPs) are essential for medication absorption and endogenous chemical efflux. Experimental screening is currently used to identify these transporter inhibitors. Because there aren't enough experimental three-dimensional protein structures, virtual screening is still difficult. An outline of the process for finding OATP2B1 transporter inhibitors in the DrugBank library of more than 5,000 pharmaceuticals and drug-like compounds is explained. The OATP member 2B1

transporter is abundantly expressed in the intestine and takes a role in the absorption of medications taken orally [43].

The role of *in silico* and *in vitro* modelling for the intestinal transport of thyrotropin-releasing hormone (TRH) analogues through PepT1 is discussed in the chapter [44]. In order to determine how structural changes affect the PepT1-mediated transport of TRH analogues, the current study uses molecular docking, molecular dynamics (MD) simulation studies, and a Caco-2 cell monolayer permeability assay. Using a homology model of the human PepT1, four TRH analogues were molecularly docked, and then the following MD simulation studies were conducted. Four TRH analogues were subjected to apical to basolateral and basolateral to apical tests on the permeability of the Caco-2 cell monolayer. Gly-Sar, a common PepT1 substrate, was used in inhibition tests to verify the PepT1-mediated transport mechanism of TRH analogues. According to MD simulation studies, the majority of substrate binding is caused by polar interactions with amino acid residues in the active site, and a decline in substrate binding was seen as bulkiness at the N-histidyl moiety of TRH analogues increased [44].

10.2 Materials and Methods

10.2.1 Experimental Data Curation

Purpose membrane transporters mediate many biological effects of chemicals and play a major role in pharmacokinetics and drug resistance. The selection of viable drug candidates among biologically active compounds requires the assessment of their transporter interaction profiles. Dataset on 3199 compositions of compounds which are potential transporters is extracted from the literature [7]. These were represented by quasi-SMILES containing data on molecular structure together with special codes related to activity in different directions (Table 10.1). The transporter behaviour data of the inhibitors were classified as the two main classes of inhibitors [versus non-inhibitors]. The data on the inhibitory activity of these 3199 compounds were assigned “1” for active and “- 1” for inactive or non-inhibitors [7]. These contain 1548 active quasi-SMILES (represented inhibitors of different quality) and 1651 inactive samples.

An example of building up a quasi-SMILES:

1. SMILES=“N1C(=NC(=C2C=1N(C=N2)[C@@H]3C[C@@H](C=C3)CO)NC4CC4)N”;
2. Code for transporter (Table 1)=[ASBT];
3. Quasi-SMILES=“N1C(=NC(=C2C=1N(C=N2)[C@@H]3C[C@@H](C=C3)CO)NC4CC4)N[ASBT]”.

Table 10.1 Quasi-SMILES code of the inhibitors

Code for quasi-SMILES	Comment
ASBT	Apical sodium-dependent bile acid transporter
BCRP	Breast cancer resistance protein
MCT1	Monocarboxylate transporter I
MDR1	Multidrug resistance protein I
MRP1	Multidrug resistance-associated protein 1–4
OATP2B1	Organic anion transporting polypeptide 2B1
OCT1	Organic cation transporter 1
PEPT1	Peptide transporter 1

10.2.2 Development of the Models

A classification-based model to forecast the inhibitor or non-inhibitor of the combined potential transporters. The so-called quasi-simplified molecular input-line entry system (quasi-SMILES), which is equivalent to the conventional SMILES, is used in QSPR/QSAR evaluations but uses all available data (not just information about the molecular structure). Such derived quasi-SMILES codes were used in the models to represent transporter behaviour [15–20, 45]. Further, the combined dataset ($n = 3199$) of the transporters was split into active-training set (ATS) (25%), passive-training set (PTS) (25%), calibration set (CS) (25%), and validation set (VS) (25%).

Using the technique of semi-correlation [14, 45] models for the inhibitory activity of different samples was built up.

$$y = C_0 + C_1 \times \text{DCW}(\text{quasi_SMILES}) \quad (10.3)$$

$$\text{DCW}(\text{quasi_SMILES}) = \sum \text{CW}(\text{code of quasi_SMILES}_k) \quad (10.4)$$

The codes for quasi-SMILES are calculated by the Monte Carlo optimisation procedure that provides the maximum of the target function

$$\text{TF} = R_A + R_P - 0.1 \times |R_A - R_P| + IIC \times W_{IIC} \quad (10.5)$$

R_A and R_P are correlation coefficient values for ATS and PTS, respectively. The IIC is the Index of Ideality of Correlation [46, 47]. The same Monte Carlo optimisation without the IIC gave significantly poorer predictive potential of the models.

In order to construct the classification model for the two classes of inhibitor (1) and non-inhibitor (– 1), additional statistical criteria like sensitivity, specificity, accuracy, and Matthews correlation coefficient (MCC) were also employed [48, 49]. The MCC

coefficient is mainly utilised in machine learning to evaluate the accuracy of binary classifications, and it can be applied when the classes have extremely disparate sizes [50].

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (10.6)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (10.7)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \quad (10.8)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (10.9)$$

In a confusion matrix, the combined two letters TP, TN, FP, and FN stand for the corresponding numbers of true positives, true negatives, false positives, and false negatives. MCC values vary from -1 to $+1$, with the former denoting a poor prediction that is exactly wrong, 0 denoting a prediction that is no better than random, and $+1$ denoting a complete adequating between predicted and observed values [50].

10.3 Result and Discussion

The calculation of the optimal descriptor (*DCW*), which is the key parameter to build any classification-based model, is using the CORAL software (<http://www.insilico.eu/coral>). The calculation of *DCW* for a given molecule here depends on the quasi-SMILES structure of that molecule, where the given quasi-structure is split into a number of the small structural attributes (*SA*), and the Monte Carlo optimisation calculates the correlation weights (*CWs*) for each *SA* of the quasi-SMILES. These *CWs* for each *SA* thus obtained by the above optimisation are added so that they constitute the full molecule; in this case, it is the quasi-SMILES structure. Table 10.2 lists the *CWs* obtained for each *SA* of the quasi-SMILES present in the molecules. An example of the calculation of the *DCW* for one of the molecules having the quasi-SMILES code (N1C(=NC(=C2C=1N(C=N2)[C@@H]3C[C@@H](C=C3)CO)NC4CC4)N[ASBT]) is provided in Table 10.3.

where N_{AT} , N_{PT} , and N_C are the numbers of *SA* in active-training set, in passive-training set, and calibration set, respectively (Tables 10.2 and 10.3). The data given in Tables 10.2 and 10.3 were obtained for $W_{IIC} = 0.5$, which is discussed in the subsequent section.

One of the aims of the given study was the assessment of ability of the *IIC* to improve the predictive potential of classification models. The W_{IIC} is weight of the *IIC*, i.e. coefficient ranged 0.1–0.7.

Table 10.2 Correlation weights (CW_s) for each SA of the quasi-SMILES

A_k	$CW(SA_k)$	ID	N_{AT}	N_{PT}	N_C	$DEFECT[SA_k]$
#...	0.5113	1	31	36	18	0.0005
(...(...	0.1667	2	219	220	194	0.0001
(...	0.0918	3	796	823	782	0
/...(...	0.4731	4	41	50	49	0.0002
/...	- 0.3147	5	53	65	57	0.0001
1...(...	- 0.2059	6	398	410	372	0
1...	0.8357	7	774	807	765	0
1.../...	- 5.0785	8	1	3	1	0.0009
2...(...	- 0.5258	9	402	421	372	0.0001
2...	0.3407	10	675	692	668	0
2.../...	- 2.5945	11	5	8	5	0.0004
2...1...	- 1.0360	12	59	59	55	0
3...(...	- 0.6801	13	298	307	274	0.0001
3...	0.0936	14	535	528	509	0
3.../...	4.8924	15	1	4	1	0.0012
3...1...	0.5043	16	11	5	12	0.0007
3...2...	0.8259	17	36	30	29	0.0002
4...(...	- 0.4946	18	211	217	184	0.0001
4...	0.5763	19	356	353	342	0
4.../...	0	20	0	1	0	0
4...1...	6.4339	21	6	3	7	0.0007
4...2...	- 2.9543	22	8	8	10	0.0002
4...3...	1.0088	23	27	20	35	0.0005
5...(...	0.1400	24	105	98	80	0.0002
5...	0.0189	25	209	202	166	0.0002
5.../...	5.6393	26	2	0	0	1
5...1...	- 7.4750	27	1	1	2	0.0007
5...2...	- 1.2957	28	1	4	3	0.0009
5...3...	- 0.2638	29	8	3	4	0.0008
5...4...	- 0.6366	30	9	6	5	0.0005
6...(...	- 1.0850	31	72	51	47	0.0004
6...	- 0.4438	32	94	78	65	0.0003
6.../...	0	33	0	3	0	0
6...1...	2.5367	34	1	3	0	1
6...3...	6.1258	35	2	0	1	0.0017
6...4...	0	36	0	1	0	0

(continued)

Table 10.2 (continued)

A_k	$CW(SA_k)$	ID	N_{AT}	N_{PT}	N_C	$DEFECT[SA_k]$
6...5...	2.0127	37	4	4	4	0
7...(...	- 0.7746	38	36	22	20	0.0005
7...	- 1.5993	39	46	32	28	0.0004
7...6...	- 1.1444	40	3	5	3	0.0004
8...(...	- 1.4578	41	15	8	7	0.0007
8...	- 1.4890	42	15	8	8	0.0006
8...7...	0.7743	43	2	1	0	1
9...(...	- 5.2887	44	1	1	3	0.001
9...	- 4.0709	45	3	2	3	0.0004
9...6...	0	46	0	1	3	0
9...8...	- 5.0209	47	1	0	0	1
= ...(...	0.0766	48	656	663	639	0
= ...	0.1613	49	724	757	716	0
= ...1...	0.2969	50	280	317	305	0.0001
= ...2...	0.3230	51	269	261	239	0.0001
= ...3...	1.6280	52	232	227	216	0
= ...4...	0.0092	53	131	135	131	0
= ...5...	2.8998	54	83	97	80	0.0001
= ...6...	- 0.1574	55	59	45	37	0.0004
= ...7...	- 0.8928	56	36	20	15	0.0007
= ...8...	- 0.8572	57	10	5	2	0.0012
= ...9...	7.5056	58	1	0	0	1
C...#...	- 0.3565	59	29	35	16	0.0005
C...(...	0.2899	60	785	806	767	0
C...	0.0453	61	795	820	783	0
C.../...	1.4637	62	41	52	52	0.0002
C...1...	0.5849	63	585	631	584	0
C...2...	1.0689	64	539	563	545	0
C...3...	0.4973	65	446	430	409	0.0001
C...4...	- 0.0300	66	267	287	265	0
C...5...	0.1184	67	163	165	128	0.0002
C...6...	- 0.5474	68	74	62	54	0.0002
C...7...	- 0.2580	69	43	23	19	0.0007
C...8...	- 0.8147	70	14	6	5	0.0009
C...9...	1.2824	71	2	1	3	0.0009
C... = ...	0.3848	72	553	559	527	0

(continued)

Table 10.2 (continued)

A_k	$CW(SA_k)$	ID	N_{AT}	N_{PT}	N_C	$DEFECT[SA_k]$
C...C...	0.8918	73	723	742	705	0
F...(...	- 0.5111	74	63	75	68	0.0001
F...	1.9909	75	73	83	76	0.0001
F...1...	0	76	0	1	0	0
F...2...	- 7.0879	77	1	0	1	0.0013
F...4...	0	78	0	0	3	0
F...C...	- 1.0458	79	41	46	32	0.0002
Br.(...	0.2533	80	7	2	3	0.0011
Br...	- 0.1213	81	7	2	3	0.0011
Br.0.2...	0	82	0	0	1	0
Br.0.3...	- 0.9756	83	1	1	0	1
Br.0.4...	- 1.0217	84	1	0	2	0.0017
Br.C...	0	85	0	1	0	0
I...(...	- 0.7582	86	2	3	6	0.0009
I...	- 0.1647	87	2	4	6	0.0009
I...3...	0	88	0	0	1	0
I...C...	0	89	0	1	2	0
Cl.(...	- 0.2604	90	28	26	26	0.0001
Cl...	1.3187	91	31	32	28	0.0001
Cl.0.1...	- 2.4033	92	4	2	5	0.0007
Cl.0.2...	0	93	0	1	0	0
Cl.0.3...	0	94	0	1	0	0
Cl.C...	- 2.9645	95	4	3	4	0.0003
N...#...	1.3416	96	25	32	15	0.0005
N...(...	- 0.8278	97	480	501	460	0
N...	- 0.3147	98	583	622	584	0
N.../...	- 1.9614	99	7	9	3	0.0007
N...1...	- 0.5937	100	142	142	151	0.0001
N...2...	- 0.3584	101	143	168	165	0.0001
N...3...	0.1814	102	144	137	133	0.0001
N...4...	- 0.1822	103	67	66	66	0
N...5...	0.6578	104	17	19	16	0.0001
N...6...	- 3.3376	105	8	4	10	0.0007
N...7...	0.5438	106	3	1	1	0.001
N...8...	0	107	0	1	3	0
N...9...	- 2.9121	108	1	0	0	1

(continued)

Table 10.2 (continued)

A_k	$CW(SA_k)$	ID	N_{AT}	N_{PT}	N_C	$DEFECT[SA_k]$
N... = ...	0.2142	109	177	178	186	0.0001
N...C...	0.5314	110	490	525	496	0
N...F...	0	111	0	1	0	0
N...Cl...	0	112	0	0	1	0
N...N...	- 2.3714	113	16	11	6	0.0007
O...(...	- 0.1308	114	695	705	683	0
O...	0.0593	115	732	756	736	0
O.../...	- 2.8947	116	4	3	5	0.0005
O...1...	1.0195	117	139	129	140	0.0001
O...2...	- 0.1934	118	83	87	95	0.0001
O...3...	1.0158	119	71	60	53	0.0002
O...4...	0.3474	120	36	26	37	0.0003
O...5...	- 0.2599	121	50	39	43	0.0002
O...6...	0.0207	122	10	9	4	0.0006
O...7...	3.6845	123	4	5	4	0.0002
O...8...	- 2.5514	124	1	0	2	0.0017
O... = ...	0.2706	125	610	602	622	0.0001
O...C...	0.0309	126	520	530	506	0
O...N...	1.7677	127	3	8	4	0.0008
O...O...	0	128	0	0	1	0
P...(...	- 1.0819	129	4	4	2	0.0005
P...	2.6365	130	7	7	2	0.0008
P...1...	- 1.1353	131	3	3	0	1
P... = ...	1.6753	132	6	3	2	0.0009
P...O...	- 2.4209	133	1	4	0	1
S...(...	- 0.7941	134	94	78	93	0.0002
S...	- 0.1689	135	117	111	124	0.0001
S.../...	1.6703	136	1	1	0	1
S...1...	0.4432	137	5	9	10	0.0005
S...2...	- 3.4664	138	9	21	20	0.0006
S...3...	4.2809	139	5	6	0	1
S...4...	0.0080	140	5	2	2	0.0008
S...5...	0	141	0	0	1	0
S...8...	0	142	0	0	1	0
S... = ...	0.3710	143	35	35	38	0.0001
S...C...	1.0060	144	46	47	48	0.0001

(continued)

Table 10.2 (continued)

A_k	$CW(SA_k)$	ID	N_{AT}	N_{PT}	N_C	$DEFECT[SA_k]$
S...N...	- 0.9896	145	10	7	6	0.0004
S...O...	8.5332	146	5	3	4	0.0004
\...(...	0.5120	147	37	39	26	0.0003
\...	0.7106	148	52	55	49	0
\...1...	- 2.7657	149	2	1	3	0.0009
\...3...	- 2.1106	150	2	1	0	1
\...4...	0	151	0	1	0	0
\...C...	0.9521	152	42	49	44	0.0001
\...N...	1.8119	153	12	13	9	0.0002
\...O...	- 2.9818	154	5	7	3	0.0006
[C +]...	7.4452	155	1	0	0	1
[BCRP]...	- 0.596	156	79	104	99	0.0002
[ASBT]...	- 2.3700	157	31	43	33	0.0002
[C@@H]...	- 0.0922	158	193	168	198	0.0002
[C@@]...	0.3572	159	86	81	80	0.0001
[C@H]...	0.0880	160	189	169	190	0.0001
[C@]...	0.4880	161	93	87	85	0.0001
[CH]...	- 0.8197	162	8	13	12	0.0003
[Br-]...	- 0.4078	163	4	2	1	0.0011
[Cl-]...	- 1.7000	164	1	1	0	1
[Br]...	2.1211	165	6	10	3	0.0009
[I-]...	- 4.3498	166	1	2	0	1
[Cl]...	1.0336	167	45	58	34	0.0004
^...	- 1.2665	168	9	7	2	0.001
^...2...	8.3159	169	1	0	0	1
^...3...	3.5091	170	1	0	0	1
^...4...	0	171	0	1	0	0
^...5...	0	172	0	2	0	0
^...C...	1.2156	173	6	3	1	0.0012
^...F...	1.6154	174	3	2	1	0.0008
^...Cl...	- 4.5392	175	1	0	0	1
^...O...	0	176	0	1	1	0
[H]...	- 1.7327	177	11	14	11	0.0002
[N +]...	- 1.0429	178	23	12	11	0.0006
[O-]...	- 0.5924	179	13	9	13	0.0003
[MCT1]...	3.0453	180	12	10	24	0.0008

(continued)

Table 10.2 (continued)

A_k	$CW(SA_k)$	ID	N_{AT}	N_{PT}	N_C	$DEFECT[SA_k]$
[MDR1]...	- 1.2082	181	399	409	377	0
[P +]...	0	182	0	0	1	0
[P-]...	- 2.6936	183	3	2	1	0.0008
[NH2]...	1.5537	184	1	1	0	1
[MRP1]...	- 1.3966	185	119	100	108	0.0002
[MRP2]...	- 1.4168	186	27	27	19	0.0003
[MRP3]...	- 0.2338	187	7	11	9	0.0003
[MRP4]...	4.2517	188	21	13	16	0.0004
[NH]...	0.7952	189	3	5	1	0.0011
[OATP2B1]...	1.1454	190	30	32	44	0.0004
[OCT1]...	4.1562	191	51	54	43	0.0001
[PEPT1]...	4.6009	192	26	26	12	0.0005
[N]...	- 4.6764	193	5	7	4	0.0004
[Se]...	1.1415	194	2	1	0	1
[Si]...	1.2054	195	1	1	1	0
[n +]...	0.8357	196	13	10	13	0.0003
[nH]...	2.8228	197	13	13	15	0.0002
c...(...	- 0.0761	198	209	211	216	0.0001
c...	0.0934	199	231	233	241	0.0001
c.../...	- 2.0681	200	2	5	3	0.0007
c...1...	0.2582	201	188	197	207	0.0001
c...2...	1.2260	202	148	139	157	0.0001
c...3...	0.6555	203	84	69	90	0.0003
c...4...	1.1246	204	67	54	62	0.0002
c...5...	1.8007	205	44	29	32	0.0004
c...6...	0.4491	206	12	10	11	0.0002
c...7...	- 3.5070	207	3	6	4	0.0005
c...8...	6.3699	208	1	1	0	1
c...9...	4.3633	209	1	1	0	1
c... = ...	- 2.7384	210	2	0	0	1
c...C...	0.5559	211	116	103	118	0.0002
c...Cl...	0.8779	212	2	2	1	0.0005
c...N...	1.3671	213	31	37	44	0.0003
c...O...	0.6487	214	79	73	89	0.0002
c...S...	- 2.7646	215	3	5	6	0.0006
c...λ...	- 1.0142	216	8	9	9	0.0001

(continued)

Table 10.2 (continued)

A_k	$CW(SA_k)$	ID	N_{AT}	N_{PT}	N_C	$DEFECT[SA_k]$
c...c...	0.6562	217	218	215	223	0.0001
n...(...	- 0.4899	218	23	26	33	0.0003
n...	- 0.7286	219	63	66	64	0
n...1...	- 1.0714	220	31	23	26	0.0003
n...2...	0.2175	221	11	7	18	0.0008
n...3...	0.9786	222	3	6	2	0.0009
n...4...	2.8537	223	4	1	2	0.0011
n...5...	0	224	0	2	1	0
n...C...	- 1.2624	225	4	7	2	0.0009
n...O...	0	226	0	1	0	0
n...c...	0.7512	227	45	55	44	0.0001
n...n...	- 1.8964	228	3	0	2	0.0015
o...(...	- 1.3584	229	9	5	4	0.0007
o...	- 0.5391	230	28	30	24	0.0001
o...1...	- 4.2328	231	6	10	6	0.0004
o...2...	0	232	0	2	1	0
o...3...	0.0319	233	1	3	3	0.0007
o...4...	5.3312	234	4	7	1	0.0012
o...5...	2.1397	235	3	3	0	1
o...6...	0	236	0	1	0	0
o...c...	0.2404	237	7	14	4	0.0009
o...n...	- 0.6625	238	21	15	18	0.0003
s...(...	4.9839	239	7	4	7	0.0005
s...	0.9463	240	24	22	27	0.0002
s...1...	1.5798	241	11	13	15	0.0003
s...2...	- 6.3189	242	4	1	2	0.0011
s...3...	5.9390	243	2	0	3	0.0015
s...4...	0	244	0	2	0	0
s...c...	2.7841	245	22	20	26	0.0003

So when we are changing the W_{IC} from 0.1 to 0.7, the changes in the classification parameters of the model such as sensitivity, specificity, accuracy, and Matthews correlation coefficient (MCC) for the different sets of the data (active-training, passive-training, calibration and validation sets) are given in Table 10.4.

The graphical variation of these parameters for the validation set is represented in Fig. 10.1a the variation of sensitivity versus different W_{IC} ; Fig. 10.1b the variation of specificity versus different W_{IC} ; Fig. 10.1c the variation of accuracy versus different W_{IC} ; and Fig. 10.1d the evolution of values of Matthews correlation coefficient

Table 10.3 Example of the calculation of the optimal descriptor (*DCW*) for quasi-SMILES “N1C(=NC(=C2C = 1N(C = N2)[C@@H]3C[C@@H](C = C3)CO)NC4CC4)N[ASBT]”

Structural attribute (SA)	<i>CW(SA)</i>	<i>ID</i>	<i>N_{AT}</i>	<i>N_{PT}</i>	<i>N_C</i>
N...	- 0.3147	98	583	622	584
1...	0.8357	7	774	807	765
C...	0.0453	61	795	820	783
(...	0.0918	3	796	823	782
=...	0.1613	49	724	757	716
N...	- 0.3147	98	583	622	584
C...	0.0453	61	795	820	783
(...	0.0918	3	796	823	782
=...	0.1613	49	724	757	716
C...	0.0453	61	795	820	783
2...	0.3407	10	675	692	668
C...	0.0453	61	795	820	783
= ...	0.1613	49	724	757	716
1...	0.8357	7	774	807	765
N...	- 0.3147	98	583	622	584
(...	0.0918	3	796	823	782
C...	0.0453	61	795	820	783
=...	0.1613	49	724	757	716
N...	- 0.3147	98	583	622	584
2...	0.3407	10	675	692	668
(...	0.0918	3	796	823	782
3...	0.0936	14	535	528	509
C...	0.0453	61	795	820	783
(...	0.0918	3	796	823	782
C...	0.0453	61	795	820	783
=...	0.1613	49	724	757	716
C...	0.0453	61	795	820	783
3...	0.0936	14	535	528	509
(...	0.0918	3	796	823	782
C...	0.0453	61	795	820	783
O...	0.0593	115	732	756	736
(...	0.0918	3	796	823	782
N...	- 0.3147	98	583	622	584
C...	0.0453	61	795	820	783
4...	0.5137	19	356	353	342

(continued)

Table 10.3 (continued)

Structural attribute (SA)	$CW(SA)$	ID	N_{AT}	N_{PT}	N_C
C...	0.0453	61	795	820	783
C...	0.0453	61	795	820	783
4...	0.5137	19	356	353	342
(...	0.0918	3	796	823	782
N...	- 0.3147	98	583	622	584
N...1...	- 0.6562	100	142	142	151
C...1...	0.5849	63	585	631	584
C...(...	0.2899	60	785	806	767
=...(...	0.0766	48	656	663	639
N... =...	0.2142	109	177	178	186
N...C...	0.5314	110	490	525	496
C...(...	0.2899	60	785	806	767
=...(...	0.0766	48	656	663	639
C... = ...	0.3848	72	553	559	527
C...2...	1.0689	64	539	563	545
C...2...	1.0689	64	539	563	545
C... =...	0.3848	72	553	559	527
=...1...	0.2344	50	280	317	305
N...1...	- 0.6562	100	142	142	151
N...(...	- 0.8278	97	480	501	460
C...(...	0.2899	60	785	806	767
C... =...	0.3848	72	553	559	527
N... =...	0.2142	109	177	178	186
N...2...	- 0.3584	101	143	168	165
2...(...	- 0.5258	9	402	421	372
C...3...	0.4973	65	446	430	409
C...(...	0.2899	60	785	806	767
C... =...	0.3848	72	553	559	527
C... =...	0.3848	72	553	559	527
C...3...	0.4973	65	446	430	409
3...(...	- 0.6801	13	298	307	274
C...(...	0.2899	60	785	806	767
O...C...	0.0309	126	520	530	506
O...(...	- 0.1307	114	695	705	683
N...(...	- 0.8278	97	480	501	460

(continued)

Table 10.3 (continued)

Structural attribute (SA)	$CW(SA)$	ID	N_{AT}	N_{PT}	N_C
N...C...	0.5314	110	490	525	496
C...4...	- 0.03	66	267	287	265
C...4...	- 0.03	66	267	287	265
C...C...	0.8918	73	723	742	705
C...4...	- 0.03	66	267	287	265
4...(...	- 0.5571	18	211	217	184
N...(...	- 0.8278	97	480	501	460
[C@@H]...	- 0.0922	158	193	168	198
[C@@H]...	- 0.0922	158	193	168	198
[ASBT]...	- 2.4325	157	31	43	33
DCW	4.9603				

for different W_{IC} . One can see that the maximal value of the MCC observed for $W_{IC} = 0.5$ (Fig. 10.1d). So the classification model at $W_{IC} = 0.5$ gives better sensitivity, accuracy, and Matthews correlation coefficient. However, the highest specificity could be obtained at $W_{IC} = 0.6$ (Fig. 10.1b).

Hence, the value of the $W_{IC} = 0.5$ should be applied to build up a model for inhibitor activity for potential Human Intestinal Transporters.

The outcomes of the classification-based models on the Human Intestinal Transporters are represented in Table 10.4. This table contains the statistical quality of these models. The statistical criteria are calculated as:

$$\text{Category}(\text{quasi_SMILES}) = \begin{cases} \text{active} & \text{if, } y \geq 0 \\ \text{inactive} & \text{if, } y < 0 \end{cases} \quad (10.10)$$

Using qualitative statistical validation metrics, such as sensitivity [0.7629–0.8067], specificity [0.7323–0.7626], accuracy [0.7526–0.7844], and Matthews correlation coefficient (MCC = [0.5058–0.5697]), a classification-based model that predicts the kind of combined inhibition (activator, non-activator) was verified. The CORAL classification model, which is being implemented to build these classification models, has the predictive potential.

Supplementary materials section contains the technical details on the model observed in the case $W_{ic} = 0.5$.

Table 10.4 Statistical quality of the model of inhibitory activity of Human Intestinal Transporter was observed for different values of the *IIC*

<i>W</i> _{IC} = 0.1	<i>W</i> _{IC} = 0.2	<i>W</i> _{IC} = 0.3	<i>W</i> _{IC} = 0.4
Classification.active.training set TP = 316 TN = 313 FP = 91 FN = 82 N = 802 Sensitivity = 0.7940 Specificity = 0.7748 Accuracy = 0.7843 MCC = 0.5688 Classification.passive.training set TP = 307 TN = 335 FP = 101 FN = 86 N = 829 Sensitivity = 0.7812 Specificity = 0.7683 Accuracy = 0.7744 MCC = 0.5488 Classification.calibration set TP = 278 TN = 313 FP = 102 FN = 91 N = 784 Sensitivity = 0.7534 Specificity = 0.7542 Accuracy = 0.7538 MCC = 0.5070	Classification.active.training set TP = 315 TN = 317 FP = 87 FN = 83 N = 802 Sensitivity = 0.7915 Specificity = 0.7847 Accuracy = 0.7880 MCC = 0.5761 Classification.passive.training set TP = 301 TN = 343 FP = 93 FN = 92 N = 829 Sensitivity = 0.7659 Specificity = 0.7867 Accuracy = 0.7768 MCC = 0.5525 Classification.calibration set TP = 273 TN = 318 FP = 97 FN = 96 N = 784 Sensitivity = 0.7398 Specificity = 0.7663 Accuracy = 0.7538 MCC = 0.5060	Classification.active.training set TP = 320 TN = 315 FP = 89 FN = 78 N = 802 Sensitivity = 0.8040 Specificity = 0.7797 Accuracy = 0.7918 MCC = 0.5838 Classification.passive.training set TP = 302 TN = 341 FP = 95 FN = 91 N = 829 Sensitivity = 0.7684 Specificity = 0.7821 Accuracy = 0.7756 MCC = 0.5503 Classification.calibration set TP = 274 TN = 312 FP = 103 FN = 95 N = 784 Sensitivity = 0.7425 Specificity = 0.7518 Accuracy = 0.7474 MCC = 0.4939	Classification.active.training set TP = 315 TN = 315 FP = 89 FN = 83 N = 802 Sensitivity = 0.7915 Specificity = 0.7797 Accuracy = 0.7855 MCC = 0.5712 Classification.passive.training set TP = 315 TN = 338 FP = 98 FN = 78 N = 829 Sensitivity = 0.8015 Specificity = 0.7752 Accuracy = 0.7877 MCC = 0.5760 Classification.calibration set TP = 280 TN = 318 FP = 97 FN = 89 N = 784 Sensitivity = 0.7588 Specificity = 0.7663 Accuracy = 0.7628 MCC = 0.5246
Classification.Validation set TP = 300 TN = 290 FP = 106 FN = 88 N = 784 Sensitivity = 0.7732 Specificity = 0.7323 Accuracy = 0.7526 MCC = 0.5058	Classification.Validation set TP = 296 TN = 296 FP = 100 FN = 92 N = 784 Sensitivity = 0.7629 Specificity = 0.7475 Accuracy = 0.7551 MCC = 0.5104	Classification.Validation set TP = 297 TN = 297 FP = 99 FN = 91 N = 784 Sensitivity = 0.7655 Specificity = 0.7500 Accuracy = 0.7577 MCC = 0.5155	Classification.Validation set TP = 302 TN = 300 FP = 96 FN = 86 N = 784 Sensitivity = 0.7784 Specificity = 0.7576 Accuracy = 0.7679 MCC = 0.5360

(continued)

Table 10.4 (continued)

Waic = 0.5	Waic = 0.6	Waic = 0.7
<p>Classification.active.training.set TP = 327 TN = 314 FP = 90 FN = 71 N = 802 Sensitivity = 0.8216 Specificity = 0.7772 Accuracy = 0.7993 MCC = 0.5993</p> <p>Classification.passive.training.set TP = 316 TN = 344 FP = 92 FN = 77 N = 829 Sensitivity = 0.8041 Specificity = 0.7890 Accuracy = 0.7961 MCC = 0.5923</p> <p>Classification.calibration.set TP = 286 TN = 323 FP = 92 FN = 83 N = 784 Sensitivity = 0.7751 Specificity = 0.7783 Accuracy = 0.7768 MCC = 0.5528</p> <p>Classification.Validation.set TP = 313 TN = 302 FP = 94 FN = 75 N = 784 Sensitivity = 0.8067 Specificity = 0.7626 Accuracy = 0.7844 MCC = 0.5697</p>	<p>Classification.active.training.set TP = 324 TN = 320 FP = 84 FN = 74 N = 802 Sensitivity = 0.8141 Specificity = 0.7921 Accuracy = 0.8030 MCC = 0.6062</p> <p>Classification.passive.training.set TP = 312 TN = 342 FP = 94 FN = 81 N = 829 Sensitivity = 0.7939 Specificity = 0.7844 Accuracy = 0.7889 MCC = 0.5776</p> <p>Classification.calibration.set TP = 279 TN = 325 FP = 90 FN = 90 N = 784 Sensitivity = 0.7561 Specificity = 0.7831 Accuracy = 0.7704 MCC = 0.5392</p> <p>Classification.Validation.set TP = 305 TN = 306 FP = 90 FN = 83 N = 784 Sensitivity = 0.7861 Specificity = 0.7727 Accuracy = 0.7793 MCC = 0.5588</p>	<p>Classification.active.training.set TP = 321 TN = 319 FP = 85 FN = 77 N = 802 Sensitivity = 0.8065 Specificity = 0.7896 Accuracy = 0.7980 MCC = 0.5962</p> <p>Classification.passive.training.set TP = 307 TN = 345 FP = 91 FN = 86 N = 829 Sensitivity = 0.7812 Specificity = 0.7913 Accuracy = 0.7865 MCC = 0.5721</p> <p>Classification.calibration.set TP = 283 TN = 325 FP = 90 FN = 86 N = 784 Sensitivity = 0.7669 Specificity = 0.7831 Accuracy = 0.7755 MCC = 0.5498</p> <p>Classification.Validation.set TP = 299 TN = 300 FP = 96 FN = 89 N = 784 Sensitivity = 0.7706 Specificity = 0.7576 Accuracy = 0.7640 MCC = 0.5282</p>

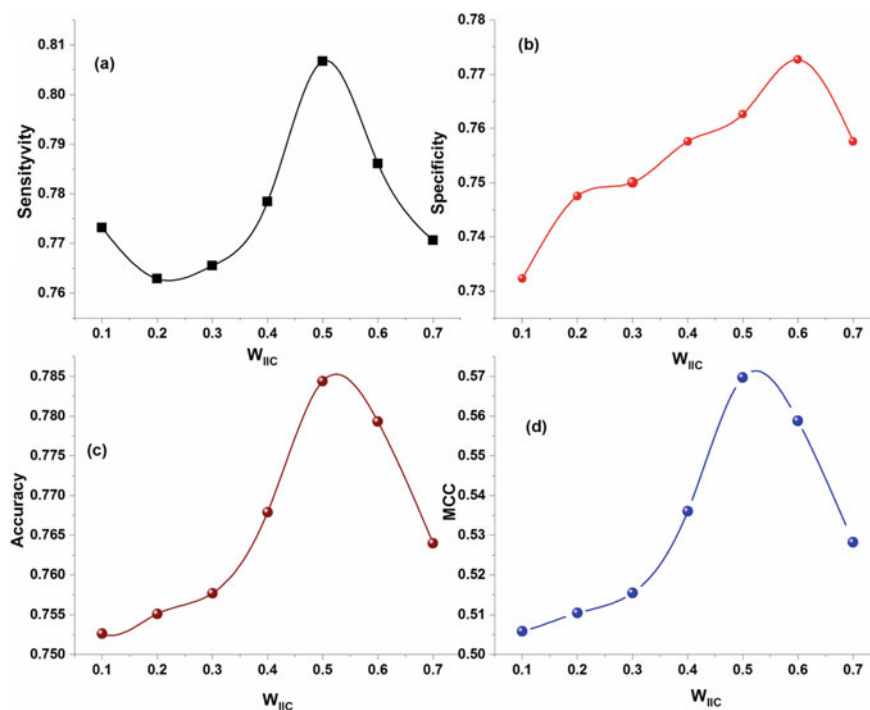


Fig. 10.1 **a** Variation of sensitivity versus different W_{IIC} , **b** the variation of specificity versus different W_{IIC} , **c** the variation of accuracy versus different W_{IIC} , and **d** the changes in Matthews correlation coefficient for different W_{IIC}

10.4 Conclusion

The classification-based QSAR model of different kinds of inhibitory activity presented in the chapter for a large list of Human Intestinal Transporter using quasi-SMILES codes was good. The extraction of biological characteristics from quasi-SMILES and computation of so-called correlation weights (CWs) for these attributes using Monte Carlo techniques proved successful in building classification-based models. As qualitative statistical validation criteria, the classification model was tested using sensitivity ($= 0.86$), specificity ($= 1$), accuracy ($= 0.96$), and Matthews correlation coefficient ($MCC = 0.90$). A model of several types of inhibitory activity using quasi-SMILES was presented for a large dataset on 3199 of the Human Intestinal Transporter. The computational experiments confirm the ability of the IIC to improve the predictive potential of classification models. So it can be said that the reported classification-based models highlighted in the present chapter are a successful attempt to predict Human Intestinal Transporters' behaviour of a large dataset. The selection of promising therapeutic candidates from libraries of bioactive

compounds should be made more accessible by understanding such features. Additionally, these profiles might be useful for modelling higher-order ADMET effects mediated by intricate transporter interactions.

Declaration of Competing Interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) *Adv Drug Deliv Rev* 23:3–25. [https://doi.org/10.1016/S0169-409X\(96\)00423-1](https://doi.org/10.1016/S0169-409X(96)00423-1)
2. Shultz MD (2018) *J Med Chem* 62:1701–1714. <https://doi.org/10.1021/acs.jmedchem.8b00686>
3. Egbert M, Whitty A, Keserü GM, Vajda S (2019) *J Med Chem* 62:10005–10025. <https://doi.org/10.1021/acs.jmedchem.8b01732>
4. Doak BC, Over B, Giordanetto F, Kihlberg J (2014) *Chem Biol* 21:1115–1142. <https://doi.org/10.1016/j.chembiol.2014.08.013>
5. DeGoey DA, Chen H-J, Cox PB, Wendt MD (2017) *J Med Chem* 61:2636–2651. <https://doi.org/10.1021/acs.jmedchem.7b00717>
6. Price E, Kalvass JC, DeGoey D, Hosmane B, Doktor S, Desino K (2021) *J Med Chem* 64:9389–9403. <https://doi.org/10.1021/acs.jmedchem.1c00669>
7. Sedykh A, Fourches D, Duan J, Hucke O, Garneau M, Zhu H, Bonneau P, Tropsha A (2013) *Pharm Res* 30:996–1007. <https://doi.org/10.1007/s11095-012-0935-x>
8. Shugarts S, Benet LZ (2009) *Pharm Res* 26:2039–2054. <https://doi.org/10.1007/s11095-009-9924-0>
9. Marquez B, Van Bambeke F (2011) *Curr Drug Targets* 12:600–620. <https://doi.org/10.2174/138945011795378504>
10. Yee SW, Chen L, Giacomini KM (2010) *Pharmacogenomics* 11:475–479. <https://doi.org/10.2217/pgs.10.22>
11. Saier Jr MH, Yen MR, Noto K, Tamang DG, Elkan C (2009) *Nucleic Acids Res* 37:D274–D278. <https://doi.org/10.1093/nar/gkn862>
12. Sarkadi B, Szakács G (2010) *Nat Rev Drug Discov* 9:897–898. <https://doi.org/10.1038/nrd3187-c1>
13. Gandhi YA, Morris ME (2009) *AAPS J* 11:541–552. <https://doi.org/10.1208/s12248-009-9132-1>
14. Vig BS, Stouch TR, Timoszyk JK, Quan Y, Wall DA, Smith RL, Faria TN (2006) *J Med Chem* 49:3636–3644. <https://doi.org/10.1021/jm0511029>
15. Toropova AP, Toropov AA (2015) *Mini Rev Med Chem* 15(2):608–621. <https://doi.org/10.2174/1389557515666150219121652>
16. Toropov AA, Toropova AP (2015) *Chemosphere* 124:40–46. <https://doi.org/10.1016/j.chemosphere.2014.10.067>
17. Achary PGR, Begum S, Toropova AP, Toropov AA (2016) *Mater Discov* 5:22–28. <https://doi.org/10.1016/j.md.2016.12.003>
18. Toropova AP, Toropov AA (2022) *Sci Total Environ* 823:153747. <https://doi.org/10.1016/j.scitotenv.2022.153747>
19. Toropova AP, Toropov AA (2021) *Int J Environ Res* 15(4):709–722. <https://doi.org/10.1007/s41742-021-00346-w>
20. Toropov AA, Kjeldsen F, Toropova AP (2022) *Chemosphere* 303:135086. <https://doi.org/10.1016/j.chemosphere.2022.135086>

21. Ta GH, Jhang C-S, Weng C-F, Leong MK (2021) *Pharmaceutics* 13:174. <https://doi.org/10.3390/pharmaceutics13020174>
22. Fang Y, Liang F, Liu K, Qaiser S, Pan S, Xu X (2018) *Food Res Int* 105:353–360. <https://doi.org/10.1016/j.foodres.2017.11.045>
23. Gonzales GB, Smagge G, Grootaert C, Zotti M, Raes K, Van CJ (2015) *Drug Metab Rev* 47:175–190. <https://doi.org/10.3109/03602532.2014.1003649>
24. Kim MT, Sedykh A, Chakravarti SK, Saiakhov RD, Zhu H (2014) *Pharm Res* 31:1002–1014. <https://doi.org/10.1007/s11095-013-1222-1>
25. Rais R, Acharya C, MacKerell Jr AD, Polli JE (2010) *Mol Pharm* 7:2240–2254. <https://doi.org/10.1021/mp100233v>
26. Zheng X, Ekins S, Raufman J-P, Polli JE (2009) *Mol Pharm* 6:1591–1603. <https://doi.org/10.1021/mp900163d>
27. Balakrishnan A, Polli JE (2006) *Mol Pharm* 3:223–230. <https://doi.org/10.1021/mp060022d>
28. Larsen S, Omkvist D, Brodin B, Nielsen C, Steffansen B, Olsen L, Jørgensen F (2009) *ChemMedChem* 4:1439–1445. <https://doi.org/10.1002/cmde.200900145>
29. Andersen R, Jørgensen FS, Olsen L, Våbenø J, Thorn K, Nielsen CU, Steffansen B (2006) *Pharm Res* 23:483–492. <https://doi.org/10.1007/s11095-006-9462-y>
30. Nielsen CU, Andersen R, Brodin B, Frokjaer S, Taub ME (2001) *Steffansen B. J Control Release* 76:129–138. [https://doi.org/10.1016/S0168-3659\(01\)00427-8](https://doi.org/10.1016/S0168-3659(01)00427-8)
31. Nielsen CU, Andersen R, Brodin B, Frokjaer S, Steffansen B (2001) *J Control Release* 73:21–30. [https://doi.org/10.1016/S0168-3659\(01\)00233-4](https://doi.org/10.1016/S0168-3659(01)00233-4)
32. Hansch C (1972) *Drug Metab Rev* 1:1–14. <https://doi.org/10.3109/03602537208993906>
33. Ekins S, Waller CL, Swaan PW, Cruciani G, Wrighton SA, Wikel JH (2000) *J Pharmacol Toxicol Methods* 44:251–272. [https://doi.org/10.1016/S1056-8719\(00\)00109-X](https://doi.org/10.1016/S1056-8719(00)00109-X)
34. Swaan PW, Koops BC, Moret EE, Tukker JJ (1998) *Recept Channels* 6:189–200. <https://doi.org/10.1023/a:1007919704457>
35. Swaan PW, Oeie S (1997) others. *J Comput Aided Mol Des* 11:581–588. <https://doi.org/10.1023/A:1007919704457>
36. Yang H, Zhai B, Wang M, Fan Y, Wang J, Cheng J, Zou J, Zhang X, Shi Y, Guo D et al. (2022) *J Ethnopharmacol* 282:114650. <https://doi.org/10.1016/j.jep.2021.114650>
37. Yamazaki S, Evers R, De Zwart L (2022) *CPT: Pharmacomet Syst Pharmacol* 11:55–67. <https://doi.org/10.1002/psp4.12733>
38. Guo K, Wang X, Huang B, Wu X, Shen S, Lin Z, Zhao J, Cai Z (2021) *Eur J Pharm Sci* 163:105839. <https://doi.org/10.1016/j.ejps.2021.105839>
39. Dechwongya P, Limpisood S, Boonnak N, Mangmool S, Takeda-Morishita M, Kulsirirat T, Rukthong P, Sathirakul K (2020) *Molecules* 25:5877. <https://doi.org/10.3390/molecules25245877>
40. Chen Y, Ma F, Jones NS, Yoshida K, Chiang P-C, Durk MR, Wright MR, Jin JY, Chinn LW (2020) *CPT: Pharmacomet Syst Pharmacol* 9:332–341. <https://doi.org/10.1002/psp4.12515>
41. Ohashi R, Watanabe R, Esaki T, Taniguchi T, Torimoto-Katori N, Watanabe T, Ogasawara Y, Takahashi T, Tsukimoto M, Mizuguchi K (2019) *Mol Pharm* 16:1851–1863. <https://doi.org/10.1021/acs.molpharmaceut.8b01143>
42. Greenhalgh K, Ramiro-Garcia J, Heinken A, Ullmann P, Bintener T, Pacheco MP, Baginska J, Shah P, Frachet A, Halder R et al (2019) *Cell Rep* 27:1621–1632. <https://doi.org/10.1016/j.celrep.2019.04.001>
43. Khuri N, Zur AA, Wittwer MB, Lin L, Yee SW, Sali A, Giacomini KM (2017) *J Chem Inf Model* 57:1402–1413. <https://doi.org/10.1021/acs.jcim.6b00720>
44. Bagul P, Khomane KS, Kesharwani SS, Pragyana P, Nandekar PP, Meena CL, Bansal AK, Jain R, Tikoo K, Sangamwar AT (2014) *J Mol Recognit* 27:609–617. <https://doi.org/10.1002/jmr.2385>
45. Toropova AP, Toropov AA (2018) *Environ Sci Pollut Res* 25:31771–31775. <https://doi.org/10.1007/s11356-018-3291-5>
46. Toropov AA, Toropova AP (2017) *Mutat Res Toxicol Environ Mutagen* 819:31–37. <https://doi.org/10.1016/J.MRGENTOX.2017.05.008>

47. Achary PGR, Toropova AP, Toropov AA (2019). Food Res Int. <https://doi.org/10.1016/j.foodres.2019.03.067>
48. A Toropov A, P Toropova A, F Rasulev B, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2012) Curr Drug Saf 7:257–261. <https://doi.org/10.2174/157488612804096542>
49. Toropova AP, Toropov AA (2017) Toxicol Lett 268:51–57. <https://doi.org/10.1016/j.toxlet.2017.01.011>
50. Dao P, Wang K, Collins C, Ester M, Lapuk A, Sahinalp SC (2011) Bioinformatics 27:i205–i213. <https://doi.org/10.1093/bioinformatics/btr245>

Chapter 11

Quasi-SMILES as a Tool for Peptide QSAR Modelling



Md. Moinul, Samima Khatun, Sk. Abdul Amin, Tarun Jha,
and Shovanlal Gayen

Abstract Peptides have played an attractive role since a few decades in the discovery of new drugs in various areas involving hormones, antimicrobials, cytokines, etc. The peptide is very righteous alternative for small molecules and biological therapeutics. Different modelling approaches can be applied to accelerate the design of different peptides-based molecules. Simplified molecular input line entry system (SMILES) is a sequence of symbols which is used to recount the molecular structure of compounds. This method helps in the development of QSAR models that describe the physicochemical property of the compounds. In contrast to SMILES, quasi-SMILES is used as an encipher for both information about molecular structure and specific experimental conditions (biological and physicochemical conditions). Quasi-SMILES uses eclectic information to design an extended representation of data. It represents all peptides in abbreviation of their corresponding amino acid and can be applied in the field of peptide-based QSAR modelling. In this chapter, we have discussed the different modelling approaches including quasi-SMILES approach for the development of QSAR models of peptide. The different models and their success in peptide QSAR models have been covered in detail.

Keywords Quasi-SMILES · SMILES · Peptide QSAR

Md. Moinul · S. Khatun · S. Gayen (✉)

Laboratory of Drug Design and Discovery, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, West Bengal 700032, India
e-mail: sgayen.pharmacy@jadavpuruniversity.in

Sk. Abdul Amin · T. Jha

Natural Science Laboratory, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, West Bengal 700032, India
e-mail: tarun.jha@jadavpuruniversity.in

Sk. Abdul Amin

Department of Pharmaceutical Technology, JIS University, 81, Nilgunj Road, Agarpara, Kolkata, West Bengal 700109, India

11.1 Introduction

Peptides are a chain of amino acids (usually 2–20). Peptide is attracting wide attention due to its high activity and selectivity with few side effects against different targets [1–3]. To date, a variety of functional peptides have been reported, including antihypertensive, antithrombotic, opioid, antimicrobial, antioxidant, anticancer, and immunomodulatory peptides [3, 4]. Thus, peptides are playing a pivotal role in drug discovery, development of vaccines, hormones, antibiotics, cytokines, neurotransmitters, immunomodulating agents, toxins, exogenous antigens, and food additives (Fig. 11.1). In comparison with small-molecule inhibitors, peptides as drug candidates have the potential to combine the properties of easy modification, remarkable specificity, excellent biocompatibility, and low side effects [3, 5].

The successful applications of peptides in drug discovery were initiated with the use of insulin in type I diabetics which was extracted from the animal pancreas. Short peptides such as oxytocin, gonadotropin-releasing hormone (GnRH), vasopressin, and somatostatin have initiated the field of peptide drug development [5]. To date, over 60 peptide drugs have been approved in the United States, Europe, and Japan to date. More than 150 peptide drugs are in the clinical development phase, and another 260 have been tested in human clinical trials [6].

Further, optimization of natural sequences of these peptides has led to the development of a number of naturally occurring hormone-mimetic peptide drugs [7]. For instance, the development anti-T2DM peptide drugs such as liraglutide, dulaglutide, and semaglutide, peptide drugs derived from GnRH such as degarelix and leuprolide and some other approved peptide drugs such as octreotide (a somatostatin mimicking

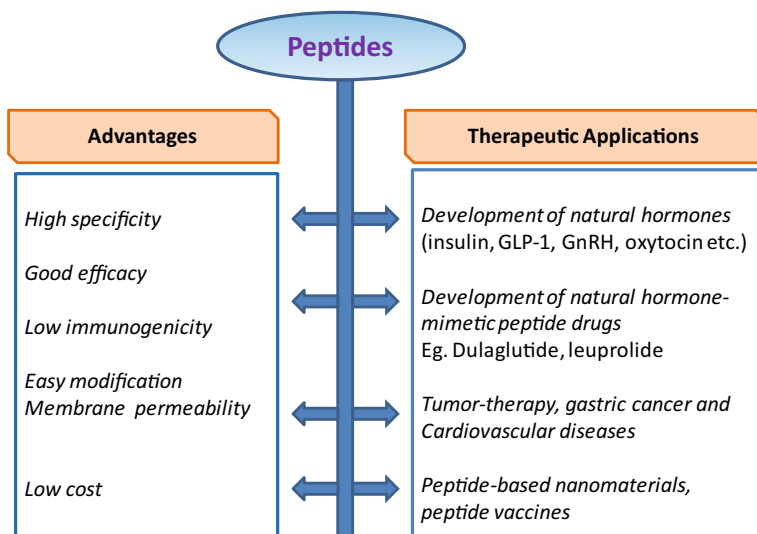


Fig. 11.1 Current applications and advantages of peptides in therapeutics

peptide drug), desmopressin (synthetic analogue of 8-Arg-vasopressin), carbetocin (an oxytocin homologue), and atosiban (an oxytocin antagonist) [5].

Peptides are also used in development of antimicrobial drugs. These antimicrobial peptides are also useful in the cosmetic industry. Antiviral peptides [8–10] drew a lot of interest during the COVID-19 pandemic. Scientists have devoted extensive effort to develop peptide vaccines against SARS-CoV-2. Design and identification of potential peptide vaccine candidates have been accelerated by the rapid application of novel technologies such as immunoinformatics analysis, *in silico* identification, epitope-based design, and molecular docking. Although, no antiviral peptide vaccine has been approved for COVID-19 treatment, and significant expertise has been gained in the development of antiviral peptide vaccines against potential future viruses, such as SARS-CoV-2 [5]. Due to their tiny size, strong affinity, ease of modification, and minimal immunogenicity, peptides have also gained interest in the treatment and diagnosis of tumours. Some altered peptides have also shown to be stable. For instance, stable-helical peptides were developed by Carvajal et al. as MDMX and MDM2 inhibitors for p53-dependent cancer treatment [11, 12]. Peptides have also demonstrated potential for treating gastric cancer. Additionally, it has been demonstrated that peptides regulate gastrointestinal (GI) motility. By boosting CGRP and endogenous PGs instead of NO, GLP-2 peripheral injection improved GI blood flow and mucosal blood flow of stomach [13].

Several peptides are identified from natural products. Some bioactive peptides obtained from plants, animals, bacteria, and fungi exhibit therapeutic properties. For example, venom peptides extracted from scorpions and snakes have been transformed for therapeutic purposes. Snake venom is believed to be a vascular endothelial growth factor (VEGF) analogue (also known as svVEGF or VEGF-F) [14–16]. Additionally, ziconotide derived from *Conus magus* venom and exenatide (a GLP-1 agonist) derived from *Gila monster* venom have both been used in the treatment of chronic neuropathic pain [17, 18]. Furthermore, another type of peptide obtained from natural products is non-ribosomal peptide (NRP). Vancomycin, lugdunin, teixobactin, and cyclosporin are antibacterial NRPs derived from bacteria and fungi, whereas amanitin, actinomycin, and nanocystin A are anti-tumour NRPs [19–22]. Some cyclodepsipeptides [23–25] (a type of NRP found in plants), such as enniatin B and emodepside [26, 27], have improved plasma stability, allowing for oral administration. Recently, recombinant technology is also employed for the longer peptide for lead discovery [5]. Peptide represents different physiological functions like natural biological messenger in endocrine signalling pathway.

Currently, *in silico* methods such as molecular docking and simulations, mathematical modelling, chemometrics, and quantum-chemical calculations are progressively being employed to design, screen, and discover bioactive peptides [28]. The QSAR modelling techniques of peptides have attracted attention in the recent years [29–31]. QSAR modelling has been extensively used to predict the physicochemical properties or biological activity of chemicals and pharmaceuticals. Designing and screening new molecules, predicting their activities, and figuring out the mechanism of bioactive peptides have all been accomplished using QSAR techniques. A variety

of techniques derived from QSAR have surfaced in the recent years [32–40]. To rationally research, evaluate, and design bioactive peptides or peptidic molecules with *in silico* assistance, computational peptidology has appeared as a distinct and promising area [41]. However, unfortunately, only few databases of peptides like CAMPR3 [42]; DBAASP [43]; BACTIBASE [44]; and CS-AMPPred [45] are available. Therefore, development of new mathematical models involving different activities of peptides is very much necessary along with conventional development of peptide containing medicines or therapy.

11.2 A Brief Overview of QSAR

The quantitative structure–property/activity relationships (QSPRs/QSARs) are a relatively emerging field in drug discovery [46]. The QSPRs/QSARs method is linked with a broad number of goals, the most important of which are likely the estimation of the physicochemical behaviour of various substances and their subsequent effect in human and animal bodies, prediction of the biochemical behaviour of various substances in medicinal aspects, and selection of substances that could be potential contender for the specific role [47]. The QSAR/QSPR approaches are based on the idea that a particular chemical compound's activity or property such as a drug binding to receptors or poisonous effect relates to its structure through a particular mathematical equation. A chemical compounds molecular structure will be related to its properties or biological activity. The prediction, interpretation, and evaluation of novel compounds with desired activities or qualities can therefore be done using this connection, lowering and simplifying the time, effort, and expense of synthesis as well as the cost of developing new products [48]. The establishment of a mathematical relationship between a chemical reaction and quantitative chemical characteristics characterizing the characteristics of the examined molecules is known as QSAR modelling on a group of structurally related chemicals. Therefore, this work aims to develop a mathematical formalism between a chemical's behaviour, or reaction, and a collection of quantitative chemical properties that may be derived from chemical structures using the appropriate experimental or theoretical methods.

Therefore, QSAR technique can be mathematically represented as

$$\text{Biological activity} = f(\text{Chemical attributes}) \quad (11.1)$$

The fundamental idea behind the term “chemical attribute” is to refer to the characteristics that specify how a chemical compound behaves, or responds [49].

11.3 Peptide QSAR Modelling

Peptide QSAR modelling involves several steps, such as dataset collection, structural characterization, variable selection, model building, model validation, and evaluation [50]. Figure 11.2 depicts a workflow of QSAR modelling of peptides.

The first important step in QSAR modelling of peptides is the dataset collection. The scope, application, and predictive power of a QSAR model depend largely on the selected dataset. The datasets can be obtained from databases, experimental results, and literature. Dearden et al. recommended avoiding datasets generated from different sources or datasets that were established using different protocols because they frequently produce unreliable modelling results [38]. The modelling results will suffer if the dataset contains duplicate samples or if two peptides have identical sequences but different endpoint values. Data collection should consider the subsequent modelling as one of the most important steps. For instance, the balance of sample size between the positive and negative groups should be taken into consideration to prevent over fitting in QSAR modelling. One of the crucial components of QSAR modelling is the characterization of molecular structures. To

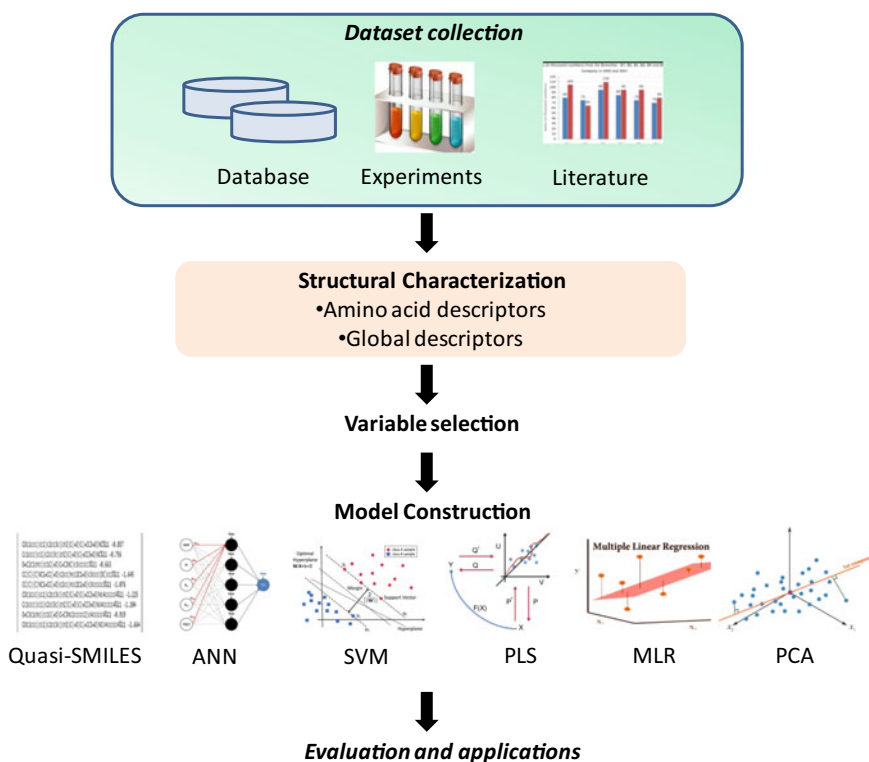


Fig. 11.2 Workflow of QSAR modelling of peptides

describe the structure of peptides, global and local descriptors—also referred to as amino acid descriptors—are frequently used. Global descriptors are molecular terms that describe an entire compound. For instance, global descriptors of molecules are those that describe a compound holistically, such as volume and polar surface area. Researchers have used global descriptors such as hERG channel inhibitors [51], chemical reactivity properties, and bioactivity scores [52] to computationally predict the potential of compounds. Some programmes such as ADMETSar [53] and ADMETLab [54] can employ global descriptors for prediction of bioactive peptides. These programmes utilize the structural characteristics of compounds annotated in SMILES code rather than amino acid sequences. The basic idea behind amino acid descriptors is to transform the amino acid sequence into a matrix–vector of structural descriptors by describing the peptide residues quantitatively. “Z-scales” (scales of hydrophilicity and bulk and electronic properties) are a set of descriptors used in peptide QSAR modelling. It is based on 29 physicochemical variables of 20 coded amino acids and is determined by principal component analysis (PCA) [55]. Later, Sandberg et al. used 5z-scales [56], which combine 26 physicochemical variables with steric, lipophilic, electronic, and other properties derived from PCA, to characterize the structures of 87 amino acids. Isotropic surface area (ISA) of the amino acid side chain and the electronic charge index (ECI) of all the atoms in the side chain are also used to interpret the peptide QSAR [57]. Moreover, peptides are 3D molecules with distinctive structures. From this viewpoint, the structural description of peptides should fairly represent their 3D properties. In this regard, global descriptors are superior to amino acid descriptors.

After structural characterization, the next important step of peptide QSAR modelling is variable selection. To guarantee the reliability and appropriate interpretation of a QSAR model, variable selection is essential. Currently, there are a lot of variable selection techniques used in QSAR modelling [58]. Some of the representative methods for variable selection include the genetic algorithm (GA), the stepwise method, forward selection, and backward elimination. Forward selection is also known as “in but not out” algorithm. In forward selection, a variable with a significant effect on dependent variables will be introduced until a new variable cannot be introduced. Backward elimination is an “out but not in” algorithm in which each variable that has no significant impact on the dependent variables is eliminated until none of the independent variables can be eliminated. The stepwise method performs forward selection and backward elimination at the same time [59], making it an efficient method for locating the optimal subspace. Genetic algorithm is a variable selection method that mimics natural selection and the natural genetic mechanism of biology [60].

After preparing the selected variables as independent variables and the correct response values of the dataset as dependent variables, the next step is to use scientific methods to build the model. This step is known as model construction. The various approaches used for peptide modelling include simplified molecular input line entry system (SMILES) and quasi-SMILES approaches [41], linear approaches like partial least square (PLS) method, multiple linear regression (MLR), and nonlinear

approaches, such as artificial neural network (ANN) and support vector machine (SVM).

Simplified molecular input line entry system (SMILES) is a specific type of chemical language or information system for defining chemical structure in a simpler way by using line notation [61]. This molecular representation can be trained faster and during training set generation improve the model and give less over fit. The molecular generation system in SMILES follows two steps, scaffold generator and decorators. Moreover, SMILES syntax is extended with aster marks [“*”]. To describe data that includes not just molecular structure but also physicochemical and/or biochemical circumstances, new expanded forms of representation must be found due to the diversity of substances used to decide activities in medicinal chemistry. Quasi-SMILES is alternative of SMILES to design the extended representation of data, which have all available eclectic information. Quasi-SMILES departs from regular simplified molecular input line entry system (SMILES) by incorporating additional symbols that encode for experiment circumstances. SMILES descriptors can be used to construct quantitative structure–property/activity relationships (QSPRs/QSARs) [62–65], whilst quasi-SMILES descriptors can be used to develop quantitative models of experimental results derived under diverse situations. It is undeniable that the quasi-SMILES strategy is encouraging better communication and collaboration between experimentalists and computational researchers [41]. The most commonly used software CORAL [66] that is based on the SMILES with string symbol helps to develop QSAR of chemical structures.

In this discussion, we have mainly focussed on QSAR studies of peptides based on quasi-SMILES tool for the development of QSAR model that will help to design new peptide molecules in the discovery and development of amino acid-based therapeutics and also in the development of peptide drug discovery in the future. This QSAR modelling will also help to improve other properties of peptide during new lead discovery, such as half-life of peptide, selectivity, potency, pharmacokinetics, and pharmacodynamics property.

11.4 SMILES-Based Descriptors for QSAR Model Development

For the development of QSAR model by using SMILES notation system, a simple mathematical equation is used for describing all descriptors which is representing in Eq. 11.2.

$$\begin{aligned} DCW(T, N) = & \textit{y}CW(\text{BOND}) + \textit{z}CW(\text{ATOMPAIR}) \\ & + \textit{x}CW(\text{NOSP}) + \textit{t}CW(\text{HALO}) + \alpha \sum CW(S_k) \\ & + \beta \sum CW(\text{SS}_k) + \gamma \sum CW(\text{SSS}_k) \end{aligned} \quad (11.2)$$

Coefficients y , x , z , t , α , β , and γ can be 0 (no) or 1 (yes). When the value of a coefficient is 1, an appropriate SMILES-based descriptor is used in model construction. If the value is 0, an appropriate SMILES-based descriptor is discarded during model construction. T and N stand for the respective threshold value and number of epochs in this equation. By using CW, the correlation weights were expressed [67–70]. To modify descriptors, various coefficients including x , y , z , and t were employed. The global SMILES qualities are represented by NOSP, HALO, BOND, and ATOMPAIR, whereas the local smile properties are indicated by S_k , SS_k , and SSS_k .

Conventional SMILES-based QSAR methods have solved different types of problem but there have few disadvantages of these methods that is why not able to solve all task specially related to peptides for development of QSAR model. This is due to the fact that in general, very complicated molecular structures of peptides and related chemical compounds cannot be described by graphs or SMILES. In the peptide QSAR modelling, instead of SMILES, quasi-SMILES can be implemented.

11.5 Quasi-SMILES

Quasi-SMILES is a technique, which is initially used for representing aspects such as circumstances and conditions associated with the substance's behaviour [71–74]. In another way, the quasi-SMILES allows for the representation of situations where the examined phenomena appear to be influenced by factors other than molecular architecture, such as physicochemical (biochemical) conditions and different environmental factors (such as the presence or absence of light, concentration, and porosity) [72]. Each condition of the substances is represented by a specific code [71]. The total of the correlation weights of the codes of conditions serves as the best descriptor. The Monte Carlo approach is used to calculate the correlation weights' numerical data [74].

The main purport for traditional QSAR model is

$$\text{Endpoint} = F(\text{molecular structure}) = F(\text{SMILES}) \quad (11.3)$$

But in case of quasi-SMILES-based QSAR modelling, the equation changes because of the eclectic data that is

$$\text{Endpoint} = F(\text{All available eclectic conditions}) = F(\text{quasi-SMILES}) \quad (11.4)$$

Toropova et al. [75] reported the representation of the quasi-SMILES based on the "SMILES + Cell Code", where cell codes are like for MCF-7, Cell Code %11; for HCT-116, Cell Code %12; for A549, Cell Code %14 and for HepG2, Cell Code %13. Further for example, if the SMILES is "COc1ccc(cc1)c2cc3c(cn2)C(=O)C(=CC3O)NC" and Cell Code is "%11", then quasi-SMILES is like "COc1ccc(cc1)c2cc3c(cn2)C(=O)C(=CC3O)NC%11". In

case of peptide QSAR, the amino acid sequence can be directly used as input for the quasi-SMILES-based model development by using Monte Carlo approach.

11.5.1 Development of QSAR Model by Quasi-SMILES

For the development of QSAR model, firstly data should be collected for different literature like “cellular uptake potentials of specific cells” [73] or “cytotoxicity of different cell line” or any other biological data [75]. To get a proper QSAR model, the biological activity data is very much important. These are represented by specific way like IC_{50} value. Then, the dataset is separated into training dataset, invisible training set, calibration set, and validation sets. The special symbol in the first place of a quasi-SMILES string denotes the distribution: active training set is denoted by +, passive training set is denoted by –, calibration set is denoted by #, and validation set is denoted by *. Here, all splits are non-identical. There are different roles of these sets which are as follows: the active training set is used for the model builder whether the invisible training set acts as a model inspector (it should check that the current model is appropriate for quasi-SMILES that are not included in the active training set). The calibration set should indicate that no overtraining has occurred. On the other hand, validation set is used for the final estimation of a model’s predictive capacity [73–75].

11.5.2 Optimal Descriptor Approach

The correlation weights of these fragments are utilized to calculate appropriate descriptors for quasi-SMILES fragments. The numerical data on the correlation weights come from the Monte Carlo optimization. Monte Carlo optimization is used to maximize value of a target function. Five steps are followed for the development of the model as described below [75–77]

Step 1: Development of the quasi-SMILES of the peptides which is nothing but the amino acid sequence.

Step 2: Correlation weights $CW(S_k)$ calculation for attributes of quasi-SMILES using so-called balance of correlations. $CW(SA_k)$ is the correlation weights for the SA_k . The numerical data on the $CW(S_k)$ should provide maximal value for the target function.

Step 3: Calculation of optimal descriptors (descriptor of correlation weights) for all quasi-SMILES by the simple equation:

$$DCW(T^*, N^*) = \sum CW(S_k) \quad (11.5)$$

The correlation weights for attributes of quasi-SMILES are calculated by the Monte Carlo method together with an example of calculation of optimal descriptor with the correlation weights.

One can identify the amino acids of two classes using numerical data on correlation weights of various amino acids that were obtained in several optimization runs: (1) amino acids with stable positive correlation weights, which are promoters of increase of pIC_{50} ; and on the other hand (2) amino acids with stable negative correlation weights, which are promoters of decrease of pIC_{50} . As a result, the method provides the models' statistical mechanistic explanation.

Step 4: Then, calculation of the model by least squares method, using quasi-SMILES of the training set:

$$\text{Potential of the model} = C_0 + C_1 \times \text{DCW}(T^*, N^*) \quad (11.6)$$

where the C_0 and C_1 are the regression coefficients.

Different types of potentiality of the model can be calculated, i.e., drug loading capacity, pIC_{50} of any therapeutic agents, antimicrobial activity of peptide, and cellular uptake in specific cell.

Step 5: Further, binary classification of the model is done by using this formula

$$\text{Class} = \begin{cases} 1, & \text{if PoM} > 0 \\ -1, & \text{if PoM} \leq 0 \end{cases} \quad (11.7)$$

Step 6: Finally, check the model predictive potential. The schematic representation is given in Fig. 11.3.

Advantages of Quasi-SMILES

- (i) These approaches offer the chance to consider all variables that might have an impact on the endpoint being studied.
- (ii) In terms of the factors that support an increase or decrease in the endpoint, it ensures a transparent interpretation of the data.
- (iii) It is possible to compare the outcomes of various data splits into active training set, passive training set, calibration set, and validation set to the integrated statistical flaws of quasi-SMILES' fragments and the quasi-SMILES algorithm itself [73].

Disadvantages of Quasi-SMILES

- (a) It is impossible to construct a model from a structured training set with a limited fraction of compounds (i.e., a composition that includes the training, invisible training, and calibration sets).
- (b) It is impossible to determine the function of quasi-SMILES' attributes that are missing from the training set [73, 75].

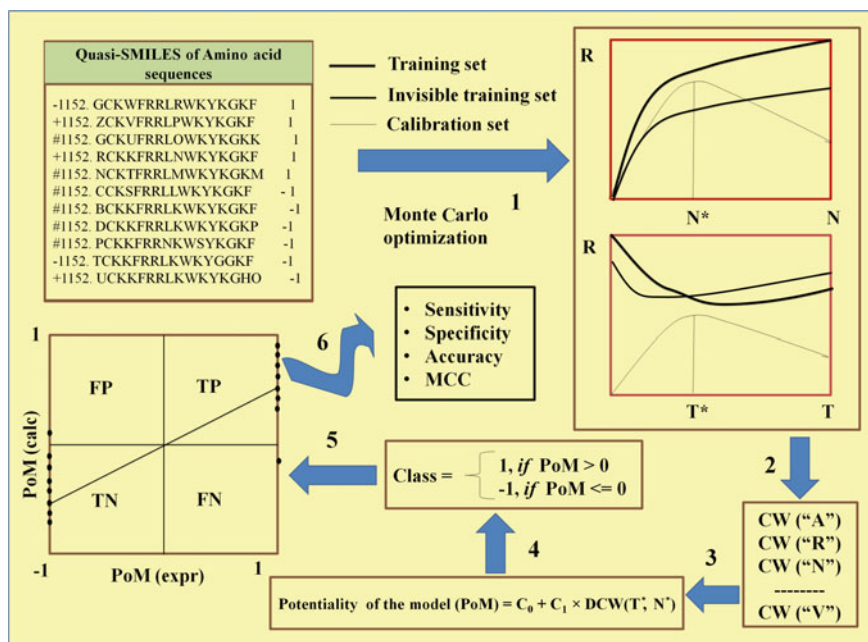


Fig. 11.3 Schematic representation of the peptide QSAR model by quasi-SMILES

11.6 Different Application of SMILES/Quasi-SMILES in Peptide QSPR/QSAR Modelling

Quasi-SMILES-based QSAR model has applications in different peptide-based QSAR modelling. Here, we have highlighted its applications mainly as antimicrobial peptides and epitope peptides with class I major histocompatibility complex (MHC).

11.6.1 Antimicrobial Peptides

In today's world, the development of novel antimicrobial peptides is very important. This is due to the fact that different bacteria are emerging as multi-drug resistant. In agricultural industry, the potent antimicrobial peptides are high in demand. As the experimental techniques for the optimization of the biological activity of the antimicrobial peptides is very time consuming as well as expensive, different computational strategies like QSAR can be applied to make the optimization process faster and cheaper. In 2015, Toropova et al. [77] established QSAR of peptides (mastoparan analogues) for their antibacterial activity. The sequence of the amino acids was used as an input for the molecular structure of the peptides. On a dataset of 33 peptides,

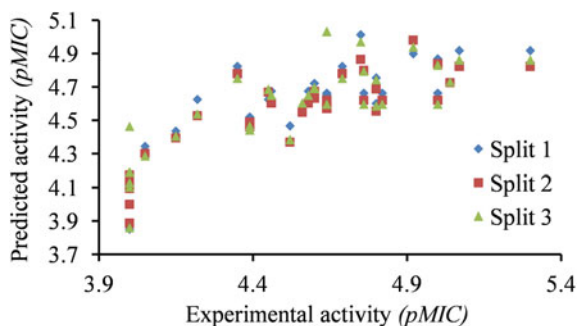
QSAR modelling was done using the best descriptors possible based on the representation of the peptide structure by the amino acid sequence. The information for the examined peptides was divided into three groups: training, calibration, and test sets. To calculate QSAR models, the Monte Carlo approach was employed as a computational tool. The definition of correlation weights was done in the beginning to get the highest value for the correlation coefficient for the calibration set. In the second step, the model was validated by using external validation set. For the external validation set, the statistical quality of QSAR for peptide antibacterial activity was as follows: $n = 7$, $r^2 = 0.8067$, $s = 0.248$ (split 1); $n = 6$, $r^2 = 0.8319$, $s = 0.169$ (split 2); and $n = 6$, $r^2 = 0.6996$, $s = 0.297$ (split 3). Other statistical parameters for the training set and calibration set of the QSAR model are shown in Table 11.1. The graphical representation of the observed and predicted values of the generated QSAR equation for different splits is shown in Fig. 11.4.

Comparing the given QSAR models to the other QSAR models developed by using 2D and 3D descriptor-based ones, the statistical parameters are better in the current QSAR models. Moreover, QSAR model generated by 3D descriptor needs high computation power and complex calculations. The QSAR study indicates that Alanine (A), Aspartic Acid (D), Phenylalanine (F), Isoleucine (I), and other amino acids can raise the $pMIC$ (negative decimal logarithm of minimum inhibitory concentrations) value. Glutamic acid (E) and serine are two amino acids that may lower the $pMIC$ value (S). Glycine (G) plays an unspecified function. Thus, the QSAR

Table 11.1 Statistical parameters for training set and calibration set of the QSAR model in case of the peptides (mastoparan analogues) for their antibacterial activity

	Number of peptides	R^2	S	Q^2	F
Training set (Split 1)	21	0.6063	0.219	0.5162	29
Training set (Split 2)	22	0.6763	0.202	0.6255	42
Training set (Split 3)	20	0.6161	0.228	0.5391	59
Calibration set (Split 1)	5	0.9678	0.108		
Calibration set (Split 2)	5	0.9630	0.278		
Calibration set (Split 3)	8	0.6819	0.222		

Fig. 11.4 Experimental versus predicted antibacterial activities of the peptides (mastoparan analogues)



modelling analysis by using Monte Carlo method by using the sequence of amino acids as input of the molecular structure can generate statistically significant QSAR models, and the generated models can be used also for the design of better active antimicrobial peptides.

In 2018, Toropova et al. [41] built a classification-based model by examining the amino acid sequences in peptides to predict the antibacterial activities of 1581 peptides that are represented by quasi-SMILES. The large set of the peptides are taken from the literature [78] and are classified as actives and inactives. A semi-correlation-based approach was used to build up models between different classes [41]. Firstly, all peptides were divided into four set which were training set, invisible training, calibration, and finally validation sets. In this case also, amino acid sequences were used as a descriptor for model building. The model was generated by using Monte Carlo optimization technique by using CORAL software. When it comes to the training, invisible training, calibration, and validation sets, the predictive potential of binary classification for antimicrobial activity for various splits was fairly strong. The statistical requirements were (i) sensitivity 0.82–0.97; (ii) specificity 0.88–0.99; (iii) accuracy 0.87–0.98; and (iv) Matthew's correlation coefficient 0.73–0.97 for the external validation sets. A plot of different statistical parameters was shown in Fig. 11.5. From Fig. 11.5, it is evident that classification-based models were fairly strong. True positive, true negative, false positive, and false negative values of different classification-based QSAR models were shown in Table 11.2. The obtained models have given insight about mechanistic insights about the biological activity of antimicrobial peptides. Attributes of the quasi-SMILES-related promoters of increase and decrease of antibacterial activity were obtained. These attributes for peptides' amino acid composition can be used to guide the design of peptides with higher antibacterial effectiveness.

Fig. 11.5 Sensitivity, specificity, accuracy, and MCC values of different classification-based QSAR models in case of antimicrobial peptides

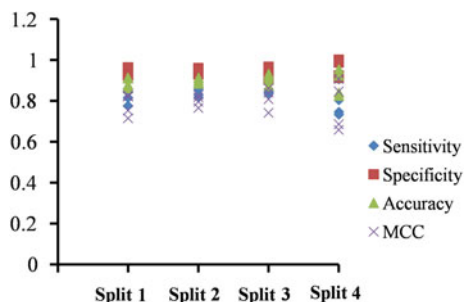


Table 11.2 True positive, true negative, false positive, and false negative values of different classification-based QSAR models in case of antimicrobial peptides

	Splits	True positive	True negative	False positive	False negative
Training set	1	146	216	10	24
	2	150	207	16	31
	3	147	213	11	27
	4	138	194	16	47
Invisible training set	1	154	205	8	28
	2	134	214	9	30
	3	172	204	12	21
	4	140	186	18	51
Calibration set	1	131	217	18	38
	2	156	206	10	24
	3	139	222	8	18
	4	123	242	0	30
Validation set	1	137	202	16	31
	2	140	219	11	24
	3	138	200	22	27
	4	145	233	3	15

11.6.2 Epitope Peptides with Class I Major Histocompatibility Complex (MHC)

Identification of epitope peptides to induce cytotoxic T lymphocytes is very important for our immune system, and it is also very important for the development of vaccines as well as immunotherapy directed against different pathogens. Major histocompatibility complex (MHC) is very important to present these peptides to T lymphocytes [79]. Thus, peptide interaction to MHC molecule is a very important step in the immunity process. The amino acid sequence can dictate the biochemical interaction between MHC-peptide complexes, and therefore, different modelling approaches can be applied to accurately predict the sequence of the peptide. In 2021, Toropova et al. [80] reported the sequence of amino acids as the basis for the development of biological activity model of these kinds of peptides. The quantitative information on class I major histocompatibility complex (MHC) molecules' biological activity with epitope peptides was collected and was randomly distributed into the active training set (25%), passive training set (25%), calibration set (25%), and validation set (25%). These different sets have different purpose in model development. Calculation of optimal correlation weights was done by the active training set, and finally, model predictive power was calculated by the validation set. The QSAR models were developed with Monte Carlo optimization with target functions TF_1 and TF_2 . The

different statistical parameters obtained from the Monte Carlo-based QSAR models were shown in Table 11.3.

From Table 11.3, it is clear that target function TF_2 may be the best approach as there are better statistical parameters observed in case of calibration set and validation set. A comparison of statistical parameters of TF_1 and TF_2 approaches for the active training set as well as validation set was shown in Fig. 11.6.

These QSAR models identify the amino acid as promoters of increase and promoters of decrease the binding affinity with MHC. Developed QSAR model showed that the amino acids like valine, leucine, phenylalanine and isoleucine,

Table 11.3 Statistical parameters of different models on epitope peptides with class I MHC

		R^2	Q^2	IIC	RMSE
<i>Active training set</i>					
Optimization with TF_1	Split 1	0.7625	0.5558	0.8732	0.36
	Split 2	0.8205	0.7052	0.9058	0.333
	Split 3	0.8846	0.8229	0.9406	0.265
Optimization with TF_2	Split 1	0.6416	0.3506	0.534	0.442
	Split 2	0.6976	0.4905	0.5568	0.432
	Split 3	0.5326	0.1846	0.7298	0.533
<i>Passive training set</i>					
Optimization with TF_1	Split 1	0.825	0.7065	0.6739	0.395
	Split 2	0.9165	0.8301	0.4709	0.374
	Split 3	0.7283	0.5982	0.8264	0.599
Optimization with TF_2	Split 1	0.7231	0.5868	0.412	0.507
	Split 2	0.9543	0.9192	0.8516	0.332
	Split 3	0.8128	0.6796	0.6251	0.562
<i>Calibration set</i>					
Optimization with TF_1	Split 1	0.6012	0.4017	0.3695	0.506
	Split 2	0.5223	0.2836	0.4258	0.592
	Split 3	0.5053	0.2612	0.3745	0.927
Optimization with TF_2	Split 1	0.9486	0.9157	0.9679	0.142
	Split 2	0.7102	0.5447	0.8406	0.337
	Split 3	0.8743	0.8139	0.8827	0.214
<i>Validation set</i>					
Optimization with TF_1	Split 1	0.622	0.4816		0.49
	Split 2	0.5481	0.3476		0.515
	Split 3	0.59	0.3277		0.7
Optimization with TF_2	Split 1	0.7766	0.6298		0.306
	Split 2	0.7856	0.6596		0.27
	Split 3	0.7909	0.6721		0.248

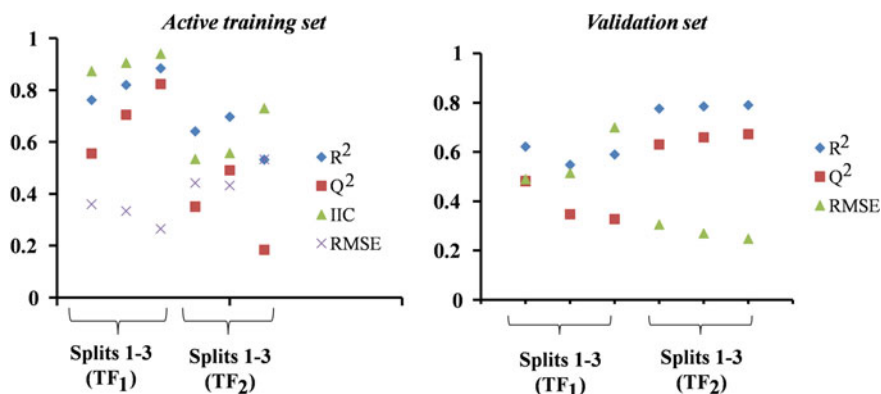


Fig. 11.6 Comparison of statistical parameters of TF₁ and TF₂ approaches for the active training set and validation set

alanine, glycine, tyrosine, etc., can increase the binding affinity value, and on the other hand, threonine and glutamic acid can decrease the binding affinity value (pIC_{50}). Thus, in this example also, simple amino acid sequence can be nicely used to develop QSAR models by using Monte Carlo approach.

11.7 Mathematical Approaches Used for Peptide QSAR Modelling

11.7.1 Multiple Linear Regressions (MLR)

The optimal QSAR model can be derived using multiple linear regression (MLR), a common mathematical modelling technique to gain more in-depth understanding of the structure–activity correlations between the chemical structure and bioactivity. MLR has the advantage of being a straightforward mathematical expression with an understandable form [18]. Despite its effective use, MLR is susceptible to descriptors that are correlated, making it unable to determine which correlated sets may be more important to the model. The best multiple linear regression (BMLR), the genetic algorithm-based multiple linear regression (GA-MLR), the heuristic method (HM), the stepwise MLR, the factor analysis MLR, and others are some of these techniques that are used recently for development of peptide QSAR [81]. Tong et al. [82] reported peptide quantitative structure activity relationship (QSAR) by using novel descriptor of amino acids (SVGER). Here, mainly amino acid descriptors were used instead of entire peptide sequences to represent the amino acid structure characteristics. It was used in two peptides, a dipeptide with a threshold of bitter taste and inhibitors of the angiotensin converting enzyme. Using stepwise multiple regression-multiple linear regression (SMR-MLR) and stepwise multiple regression-partial least square

regression (SMR-PLS), QSAR models were created. Coefficient of correlation R_{cum}^2 was employed to estimate how well the model fit the data. The model was based on the correlation coefficient between cross-validation and observed activities (Q_{LOO}^2) for internal validation and Q_{ext}^2 for external validation.

Masand et al. [83] built a peptide QSAR model for finding out the special structural feature in peptide type of inhibitors responsible for the SARS-CoV inhibition by using genetic algorithm–multi-linear regression (GA-MLR) methodology with the help of QSARINS ver. 2.2.2 software.

11.7.2 *Partial Least Square (PLS)*

PLS is widely utilized in many different industries. The PLS model attempts to determine the multidimensional direction in X space that best describes the highest multidimensional variance direction in Y space [49]. The ability to interpret the influence of descriptors on output prediction is the main advantage of PLS models. PLS is well-known in the realm of QSAR/QSPR for its use with CoMFA and CoMSIA. PLS has recently changed by combining with other mathematical techniques to perform better in QSAR/QSPR analysis. There have different types of PLS like genetic partial least squares (G/PLS), orthogonal signal correction partial least squares (OSC-PLS), and factor analysis partial least squares (FA-PLS) [81]. In 2007, Jenssen et al. [84] published peptide QSAR results using Simca-P 10.0 software and PLS techniques to find out the antimicrobial activity of peptide.

11.7.3 *Principal Component Analysis (PCA)*

Principal component analysis, or PCA, is a technique for reducing the number of dimensions in large data sets by condensing a large collection of variables into a smaller set that retains the majority of the large set's information. Mahmoodi-Reihani et al. [85] developed a peptide QSAR model to calculate numerical descriptive vectors (NDVs) for peptide sequences that was based on the physicochemical properties of amino acids (AAs) and principal component analysis (PCA).

For the development of composite variables, PLS and PCA function somewhat differently. Whilst PLS builds its composite variables to explain the maximum variability in the response within the context of linear regression, PCA builds its composite variables to explain the maximum variability in all the original predictors, or the explanatory variables of interest [86].

11.7.4 Genetic Algorithm (GA)-Based Peptide QSAR

This is another type of algorithm for the design of peptide QSAR. A binary string termed a chromosome, which defines each individual in the population, represents a subset of descriptors (Fig. 11.7). There are as many genes on the chromosome as there are descriptions. If the matching descriptor is chosen in the model, a gene is given the value 1; otherwise, it is given the value 0. The initial population of chromosomes is created during GA initialization. A generation is the development of a new population from an existing one. A fitness function in each generation makes sure that only the fittest chromosomes pass on their genes to the following one. A local change in a chromosome is produced by a second procedure called mutation, which is administered with a modest chance. The fitness function and selection process, along with the crossover and mutation procedures, are necessary to generate variation within the population, which leads to learning and evolution towards an optimum solution. One distinguishing characteristic of a GA is that, in keeping with Darwinian evolution, only the fittest chromosomes are allowed to pass on their traits to the following generation [87].

Andrade-Ochoa et al. [88] applied genetic algorithm-variable subset selection for peptide QSAR model generation with MobyDigs software. To establish which structural arrangement and functional groups are most crucial for biological activity, QSAR models were only run with structural descriptors.

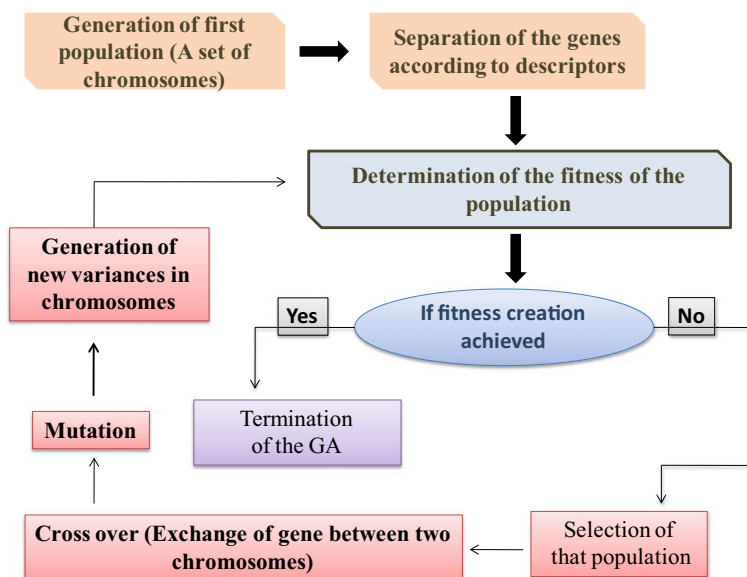


Fig. 11.7 Schematic representation of the genetic algorithm

11.7.5 Particle Swarm Optimization Algorithm (PSO)

Each individual particle in the multidimensional search space is a potential solution for the PSO algorithm. Every particle's updated location is influenced by its own and the swarm's collective experience in each generation; specifically, each particle's velocity is adjusted in the direction of its own personal best position (P_i) and the overall best position (P_g). The PSO algorithm limits each particle's position to the 0 and 1 binary search space, and the velocity denotes the likelihood that each dimension's position will take the value 1 or 0. The velocity updating equation does not change, and a sigmoid function maps each dimension's velocity to the range [0, 1]. Schematic representation of the PSO-GA-SVM scheme for peptide QSAR is depicted in Fig. 11.8.

Zhou et al. [89] proposed a novel method based on PSA-GO-SVM in order to fully utilize the advantages of genetic algorithm (GA) and particle swarm optimization (PSO) algorithm. The PSO-GA-SVM scheme is illustrated in Fig. 11.8. In this method, the kernel parameters of SVM were optimized, and the optimized features subset was simultaneously determined. In order to evaluate the proposed method, four peptide datasets were employed for the investigation of QSAR. The structural and physicochemical features of peptides from amino acid sequences were used to represent peptides for QSAR. A protein dataset of 277 proteins was employed to

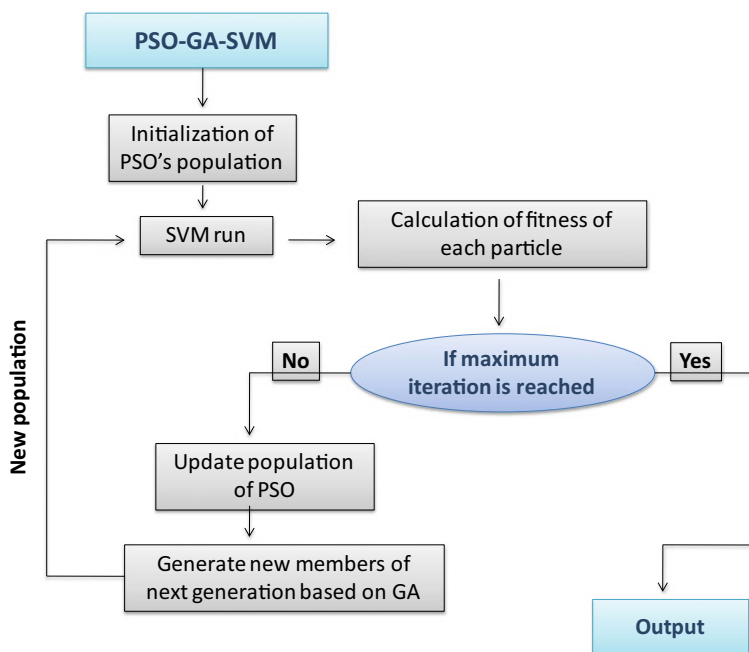


Fig. 11.8 Representation of the PSO-GA-SVM scheme for peptide QSAR

evaluate the proposed method to predict the structural class of protein. Good results were obtained which indicated that the proposed method may have a great potential for usage as a tool in peptide QSAR and protein prediction research [89].

11.7.6 Artificial Neural Network (ANN)

ANN is a type of artificial intelligence that attempts to imitate some of the qualities of neural networks. In case of antimicrobial peptide discovery, ANN is represented by a network of descriptors, which can be thought of as input nodes or neurons. These nodes are linked together to form a network, which is then transformed in a hidden layer to produce an output node (Fig. 11.9). The ability of neural networks to naturally model nonlinear systems is one of their advantages. The potential to over fit the data and the difficulty in determining which descriptors are most important in the final model are drawbacks of this method [90].

He et al. [91] built a peptide QSAR model with the help of ANN algorithm and finally designed some ACE inhibitor peptide. In order to model the neural network, seven hidden layer neurons were chosen. Repeated modelling showed that the correlation coefficient R reached 0.928, the mean square error for the training set was 0.0188, and the mean square error for the prediction set was 0.2091. This study also suggested that Alcalase was a suitable protease for the production of ACE-inhibitory peptides, and C-terminal is particularly significant to ACE-inhibitory action. Proteins

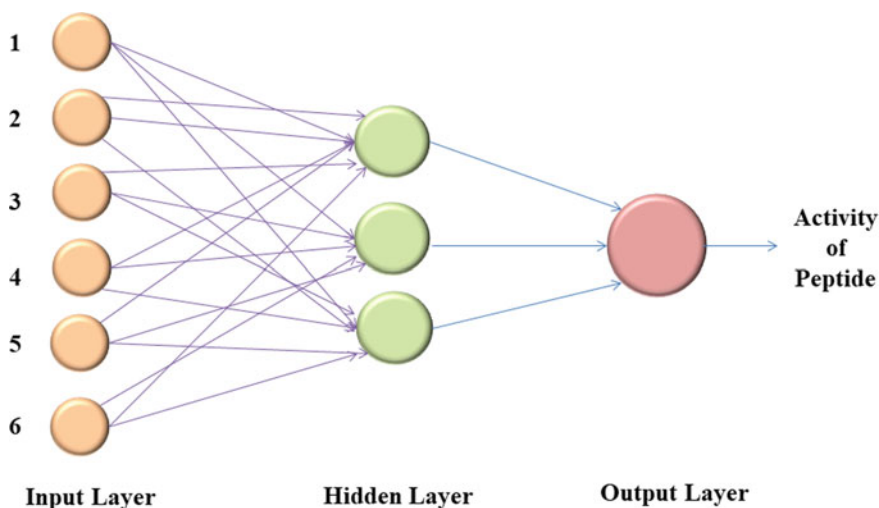
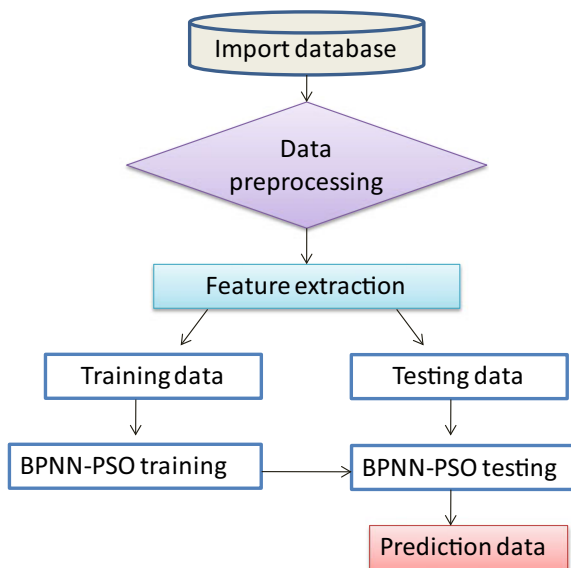


Fig. 11.9 Artificial neural network for peptide QSAR modelling. Input layer represents the descriptors of the peptide structure, the hidden layer illustrates the transformations of the input layer to a reduced level, and finally, the output layer is associated with activity of the peptide

Fig. 11.10 Schematic representation of proposed ANN-BPNN-based model



containing rich hydrophobic amino acids are also possible good sources to produce ACE-inhibitory peptides.

Rajkumar et al. [92] utilized ANN approach with a back propagation neural network (BPNN) to detect the antifungal, antibacterial, and antiviral effects of antimicrobial peptides (AMPs). In the proposed model, BPNN was used to build an ANN framework that aids in the optimal categorization of peptide sequences with antimicrobial activity (Fig. 11.10).

The BPNN is trained on the datasets, and then, a PSO algorithm was used to avoid over fitting. As a result, during testing, the BPNN clearly finds predicted samples pertaining to the same classes, avoiding the problem of false positives. The simulation is used to assess the model's efficacy against various metrics such as accuracy, precision, recall, and f1-measure. The performance of the BPNN-PSO model demonstrates its effectiveness in classifying instances faster than other techniques. The principle is simple, easy to programme, converges faster, and it generally provides a better solution [92].

11.7.7 Support Vector Machine (SVM)

SVM is essentially utilized as a classification method, with a hyperplane acting as a barrier between two classes (H). The margin between the two classes is measured by the distances between plane H and the planes cutting the closest sample points on either side of H, namely H_1 and H_2 . The optimized plane is then defined as the

one that maximizes this margin. In particular, support vectors are defined as sample points that are perfectly positioned on planes H_1 and H_2 [28].

Zhou et al. [93] reported peptide QSAR modelling for systematic comparison and comprehensive evaluation of the 80 amino acid descriptors (AADs) by using linear PLS, GA, and nonlinear SVM. 11 structural and physicochemical characteristics of peptide, including amino acid composition, dipeptide composition, autocorrelation, composition, transition and distribution, sequence order, and pseudo-amino acid composition, were used to define peptide from amino acid sequences. This research also indicated that adding more new AADs with more diverse original features would not significantly enhance their performance in peptide QSAR modelling. Instead, the AAD characterization of peptide sequences can be handled using multivariate algorithms that take into account residue interaction, context effect, and conformational factor, amongst other things.

11.7.8 Other Methods

Ant colony optimization algorithm (COA) and artificial immunization algorithm are also employed for the feature selection of any derivatives.

11.8 Conclusions

Peptides have recently emerged as a distinct class of bioactive molecules due to their high therapeutic potential. Several peptides are in the clinical development phase, and more than 80 have already made it to the market on a global scale. Peptide drugs are used to treat a variety of diseases, including cancer, cardiovascular disease, diabetes mellitus, digestive disorders, infectious diseases, and in the development of vaccines. We anticipate that therapeutic peptides will continue to draw funding and research attention due to their enormous therapeutic potential, economic value, and market potential. In silico approaches such as QSAR have been employed to identify, screen, and discover peptides. On the one hand, we need to emphasize more on the benefits of QSAR such as how it can be used to probe the mechanism(s) of action and significantly cut down on the time and expense associated with peptide identification and evaluation. On the other hand, we must confront the challenges of QSAR when applied to peptides, such as the difficulty in obtaining high-quality datasets, limited number of descriptors to generate models, and selection of model building methods which is a requirement for QSAR modelling. The method of model construction is an important factor in peptide QSAR modelling. Applications of SMILES, quasi-SMILES, machine learning algorithms, and artificial intelligence in QSAR have received enough attention in the recent past. In the current work, various model building techniques are discussed, giving special emphasis to SMILES and quasi-SMILES approaches. However, it is challenging to suggest a particular method as

the best and only method for QSAR modelling of peptides due to the differences in the quality of the chosen sample, numbers, and structural parameters. We should not only rely on established modelling techniques, but also consciously apply novel modelling techniques or incorporate integrate modelling methodologies such as the sample grouping method and parameter selection algorithm. It is important to test a variety of approaches or combination strategies to accomplish QSAR analysis. Improved mathematical methods in the quasi-SMILES construction can be helpful for better statistical quality. There are not many peptide QSAR studies already available. Therefore, extensive research is required to advance our understanding for using QSAR approach in peptide drug discovery.

References

1. Kang L, Han T, Cong H, Yu B, Shen Y (2022) *BioFactors* 8(3):575–596. <https://doi.org/10.1002/biof.1822>
2. Apostolopoulos V, Bojarska J, Chai TT, Elnagdy S, Kaczmarek K, Matsoukas J, New R, Parang K, Lopez OP, Parhiz H, Perera CO, Pickholz M, Remko M, Saviano M, Skwarczynski M, Tang Y, Wolf WM, Yoshiya T, Zabrocki J, Zielenkiewicz P, Alkhazindar M, Barriga V, Kelaidonis K, Sarasia EM, Toth I (2021) *Molecules* 26(2):430. <https://doi.org/10.3390/molecules26020430>
3. Bhat ZF, Kumar S, Bhat HF (2015) *J Food Sci Technol* 52(9):5377–5392. <https://doi.org/10.1007/s13197-015-1731-5>
4. Yosten GL, Elrick MM, Salvatori A, Stein LM, Kolar GR, Ren J, Corbett JA, Samson WK (2015) *Peptides* 72:192–195. <https://doi.org/10.1016/j.peptides.2015.05.011>
5. Wang L, Wang N, Zhang W, Cheng X, Yan Z, Shao G, Wang X, Wang R, Fu C (2022) *Signal Transduct Target Ther* 7(1):48. <https://doi.org/10.1038/s41392-022-00904-4>
6. Lau JL, Dunn MK (2018) *Bioorg Med Chem* 26(10):2700–2707. <https://doi.org/10.1016/j.bmc.2017.06.052>
7. Qvit N, Rubin SJS, Urban TJ, Mochly-Rosen D, Gross ER (2017) *Drug Discov Today* 22(2):454–492. <https://doi.org/10.1016/j.drudis.2016.11.003>
8. Malonis RJ, Lai JR, Vergnolle O (2020) *Chem Rev* 120(6):3210–3229. <https://doi.org/10.1021/acs.chemrev.9b00472>
9. Chourasia R, Padhi S, Phukon LC, Abedin MM, Sirohi R, Singh SP, Rai AK (2022) *Bioengineered* 13(4):9435–9454. <https://doi.org/10.1080/21655979.2022.2060453>
10. Chew MF, Poh KS, Poh CL (2017) *Int J Med Sci* 14(13):1342–1359. <https://doi.org/10.7150/ijms.21875>
11. Carvajal LA, Neriah DB, Senecal A, Benard L, Thiruthuvanathan V, Yatsenko T, Narayanagari SR, Wheat JC, Todorova TI, Mitchell K, Kenworthy C (2018) *Sci Transl Med* 10(436):eaao3003. <http://doi.org/10.1126/scitranslmed.aao3003>
12. Carvajal LA, Ben-Neriah D, Senecal A, Bernard L, Narayanagari SR, Kenworthy C, Thiruthuvanathan V, Guerlavais V, Annis DA, Bartholdy B, Will B (2017) *Blood* 130:795. https://doi.org/10.1182/blood.V130.Suppl_1.795.795
13. Suyen GG, Isbil-Buyukcuskun N, Cam B, Ozluk K (2015) *Peptides* 64:62–66. <https://doi.org/10.1016/j.peptides.2014.12.008>
14. Brown MC, Calvete JJ, Staniszewska I, Walsh EM, Perez-Liz G, Del Valle L, Lazarovici P, Marcinkiewicz C (2017) *Growth Factors* 25(2):108–117. <https://doi.org/10.1080/089771907.01532385>
15. Yamazaki Y, Matsunaga Y, Tokunaga Y, Obayashi S, Saito M, Morita T (2009) *J Biol Chem* 284(15):9885–9891. <https://doi.org/10.1074/jbc.M809071200>

16. Toivanen PI, Nieminen T, Laakkonen JP, Heikura T, Kaikkonen MU, Ylä-Herttua S (2017) *Sci Rep* 7(1):1. <https://doi.org/10.1038/s41598-017-05876-y>
17. Schmidtko A, Lötsch J, Freynhagen R, Geisslinger G (2010) *The Lancet* 375(9725):1569–1577. [https://doi.org/10.1016/S0140-6736\(10\)60354-6](https://doi.org/10.1016/S0140-6736(10)60354-6)
18. Pope JE, Deer TR, Amirdelfan K, McRoberts WP, Azeem N (2017) *Curr Neuropharmacol* 15(2):206–216. <http://doi.org/10.2174/1570159x14666160210142339>
19. Schilling NA, Berscheid A, Schumacher J, Saur JS, Konnerth MC, Wirtz SN, Beltrán-Beleña JM, Zipperer A, Krismer B, Peschel A, Kalbacher H (2019) *Angew Chem* 58(27):9234–9238. <https://doi.org/10.1002/anie.201901589>
20. Bitschar K, Sauer B, Focken J, Dehmer H, Moos S, Konnerth M, Schilling NA, Grond S, Kalbacher H, Kurschus FC, Götz F (2019) *Nat Commun* 10(1):1. <https://doi.org/10.1038/s41467-019-10646-7>
21. Niu X, Thaochan N, Hu Q (2020) *J Fungus* 6(2):61. <https://doi.org/10.3390/jof6020061>
22. Brown AS, Calcott MJ, Owen JG, Ackerley DF (2018) *Nat Prod Rep* 35(11):1210–1228. <https://doi.org/10.1039/c8np00036k>
23. Gould A, Ji Y, Aboye TL, Camarero JA (2011) *Cyclotides*. *Curr Pharm Des* 17(38):4294–4307. <http://doi.org/10.2174/138161211798999438>
24. Sivanathan S, Scherkenbeck J (2014) *Molecules* 19(8):12368–12420. <http://doi.org/10.3390/molecules190812368>
25. Weidmann J, Craik DJ (2016) *J Exp Bot* 67(16):4801–4812. <https://doi.org/10.1093/jxb/erw210>
26. Prosperini A, Berrada H, Ruiz MJ, Caloni F, Coccini T, Spicer LJ, Perego MC, Lafranconi A (2017) *Front Public Health* 5:304. <https://doi.org/10.3389/fpubh.2017.00304>
27. Martin RJ, Buxton SK, Neveu C, Charvet CL, Robertson AP (2012) *Exp Parasitol* 132(1):40–46. <https://doi.org/10.1016/j.exppara.2011.08.012>
28. Nongonierma AB, FitzGerald RJ (2017) *Trends Food Sci Technol* 69:289–305. <https://doi.org/10.1016/j.tifs.2017.03.003>
29. Yao XJ, Panaye A, Doucet JP, Zhang RS, Chen HF, Liu MC, Hu ZD, Fan BT (2004) *J Chem Inf Comp Sci* 44(4):1257–1266. <https://doi.org/10.1021/ci049965i>
30. Qin S, Liu H, Wang J, Yao X, Liu M, Hu Z, Fan B (2007) *QSAR Comb Sci* 26(4):443–451. <https://doi.org/10.1002/qsar.200630059>
31. Chamjangali MA, Beglari M, Bagherian G (2007) *J Mol Graph Model* 26(1):360–367. <https://doi.org/10.1016/j.jmgm.2007.01.005>
32. FitzGerald RJ, Cermeño M, Khalesi M, KleeKayai T, Amigo-Benavent M (2020) *J Funct Foods* 64:103636. <https://doi.org/10.1016/j.jff.2019.103636>
33. Pripp AH (2006) *J Agric Food Chem* 54(1):224–228. <https://doi.org/10.1021/jf0521303>
34. Pripp AH, Isaksson T, Stepaniak L, Sørhaug T, Ardö Y (2005) *Trends Food Sci Technol* 16(11):484–494. <https://doi.org/10.1016/j.tifs.2005.07.003>
35. Holton TA, Vijayakumar V, Khaldi N (2013) *Trends Food Sci Technol* 34(1):5–17. <https://doi.org/10.1016/j.tifs.2013.08.009>
36. Bączek T, Kalisz R (2009) *Proteomics* 9(4):835–847. <https://doi.org/10.1002/pmic.200800544>
37. Bahadori M, Hemmateenejad B, Yousefinejad S (2019) *Amino Acids* 51(8):1209–1220. <https://doi.org/10.1007/s00726-019-02761-y>
38. Dearden JC, Cronin MT, Kaiser KL (2009) *SAR QSAR Environ Res* 20(3–4):241–266. <https://doi.org/10.1080/10629360902949567>
39. Doytchinova IA, Flower DR (2001) *J Med Chem* 44(22):3572–3581. <https://doi.org/10.1021/jm010021j>
40. Pal R, Jana G, Sural S, Chattaraj PK (2019) *Chem Biol Drug Des* 93(6):1083–1095. <https://doi.org/10.1111/cbdd.13428>
41. Toropova AP, Toropov AA, Benfenati E, Leszczynska D, Leszczynski J (2018) *BioSystems* 169:5–12. <http://doi.org/10.1016/j.biosystems.2018.05.003>
42. Waghu FH, Barai RS, Gurung P, Idicula-Thomas S (2016) *Nucleic Acids Res* 44(D1):D1094–D1097. <https://doi.org/10.1093/nar/gkv1051>

43. Pirtskhalava M, Gabrielian A, Cruz P, Griggs HL, Squires RB, Hurt DE, Grigolava M, Chubinidze M, Gogoladze G, Vishnepolsky B, Alekseev V (2016) *Nucleic Acids Res* 44(D1):D1104–D1112. <https://doi.org/10.1093/nar/gkv1174>
44. Hammami R, Zouhir A, Ben Hamida J, Fliss I (2007) *BMC Microbiol* 7(1):1. <https://doi.org/10.1186/1471-2180-7-89>. DOI:10.1186/1471-2180-7-89
45. Porto WF, Pires AS, Franco OL (2012) *PLoS One* 7(12):e51444. <https://doi.org/10.1371/journal.pone.0051444>
46. Hasan MR, Alsaiani AA, Fakhurji BZ, Molla MH, Asseri AH, Sumon MA, Park MN, Ahammad F, Kim B (2022) *Molecules* 27(13):4169. <https://doi.org/10.3390/molecules27134169>
47. Toropov AA, Toropova AP (2020) *Molecules* 25(6):1292. <https://doi.org/10.3390/molecules25061292>
48. Chtita S, Bouachrine M, Lakhli T (2016) *Revue Interdisciplinaire* 1(1)
49. Roy K, Kar S, Das RN (2015) *Fundamental concepts*. Springer, Cham. <http://doi.org/10.1007/978-3-319-17281-1>
50. Bo W, Chen L, Qin D, Geng S, Li J, Mei H, Li B, Liang G (2021) *Trends Food Sci Technol* 114:176–188. <https://doi.org/10.1016/j.tifs.2021.05.031>
51. Sinha N, Sen (2011) *Eur J Med Chem* 46(2):618–630. <http://doi.org/10.1016/j.ejmech.2010.11.042>
52. Flores-Holguín N, Frau J, Glossman-Mitnik D (2019) *Theor Chem Acc* 138(6):1–9. <https://doi.org/10.1007/s00214-019-2469-3>
53. Yang H, Lou C, Sun L, Li J, Cai Y, Wang Z, Li W, Liu G, Tang Y (2019) *Bioinformatics* 35(6):1067–1069. <https://doi.org/10.1093/bioinformatics/bty707>
54. Dong J, Wang NN, Yao ZJ, Zhang L, Cheng Y, Ouyang D, Lu AP, Cao DS (2018) *J Cheminformatics* 10(1):1. <https://doi.org/10.1186/s13321-018-0283-x>
55. Hellberg S, Sjoestrom M, Skagerberg B, Wold S (1987) *J Med* 30(7):1126–1135. <https://doi.org/10.1021/jm00390a003>
56. Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S (1998) *J Med Chem* 41(14):2481–2491. <https://doi.org/10.1021/jm9700575>
57. Collantes ER, Dunn WJ III (1995) *J Med Chem* 38(14):2705–2713. <https://doi.org/10.1021/jm00014a022>
58. Shahlaei M (2013) *Chem Rev* 113(10):8093–8103. <https://doi.org/10.1021/cr3004339>
59. Smith DW, Gill DS, Hammond JJ (1985) *J Stat Comput Simul* 22(3–4):217–227. <https://doi.org/10.1080/00949658508810848>
60. Niazi A, Leardi R (2012) *J Chemom* 26(6):345–351. <https://doi.org/10.1002/cem.2426>
61. Weininger DJ (1988) *Chem Inf Comput Sci* 28(1):31–36. <http://doi.org/10.1021/ci00057a005>
62. Ghosh K, Amin SA, Gayen S, Jha T (2021) *J Mol Struct* 1224:129026. <https://doi.org/10.1016/j.molstruc.2020.129026>
63. Jain S, Bhardwaj B, Amin SA, Adhikari N, Jha T, Gayen S (2020) *J Biomol Struct Dyn* 38(6):1683–1696. <https://doi.org/10.1080/07391102.2019.1615000>
64. Toropov AA, Benfenati E (2007) *Comput Biol Chem* 31(1):57–60. <https://doi.org/10.1016/j.combiolchem.2007.01.003>
65. Jain S, Amin SA, Adhikari N, Jha T, Gayen S (2020) *J Biomol Struct Dyn* 38(1):66–77. <https://doi.org/10.1080/07391102.2019.1566093>
66. <http://www.insilico.eu/coral>
67. Gaikwad R, Ghorai S, Amin SA, Adhikari N, Patel T, Das K, Jha T, Gayen S (2018) *Toxicol In Vitro* 52:23–32. <https://doi.org/10.1016/j.tiv.2018.05.016>
68. Amin SA, Bhargava S, Adhikari N, Gayen S, Jha (2018) *J Biomol Struct Dyn* 36(3):590–608. <http://doi.org/10.1080/07391102.2017.1288659>
69. Veselinović JB, Nikolić GM, Trutić NV, Živković JV, Veselinović AM (2015) *SAR QSAR Environ Res* 26(6):449–460. <https://doi.org/10.1080/1062936X.2015.1049665>
70. Toropov AA, Toropova AP, Lombardo A, Roncaglioni A, Benfenati E, Gini G (2011) *Eur J Med Chem* 46(4):1400–1403. <https://doi.org/10.1016/j.ejmech.2011.01.018>
71. Toropov AA, Achary PG, Toropova AP (2016) *Chem Phys Lett* 660:107–110. <https://doi.org/10.1016/j.cplett.2016.08.018>

72. Toropova AP, Toropov AA, Leszczynska D, Leszczynski J (2017) *Ecotoxicol Environ Saf* 139:404–407. <https://doi.org/10.1016/j.ecoenv.2017.01.054>
73. Toropova AP, Toropov AA, Leszczynska D, Leszczynski J (2021) *Comput Biol* 136:104720. <https://doi.org/10.1016/j.combiomed.2021.104720>
74. Toropov AA, Toropova AP, Benfenati E, Diomedea L, Salmons M (2018) *Struct Chem* 29(4):1213–1223. <https://doi.org/10.1007/s11224-018-1115-3>
75. Toropova AP, Toropov AA (2019) *Mol Divers* 23(2):403–412. <https://doi.org/10.1007/s11030-018-9881-9>
76. Toropov AA, Rasulev BF, Leszczynski J (2008) *Bioorg Med Chem* 16(11):5999–6008. <https://doi.org/10.1016/j.bmc.2008.04.055>
77. Toropova MA, Veselinović AM, Veselinović JB, Stojanović DB, Toropov AA (2015) *Comput Biol Chem* 59:126–130. <https://doi.org/10.1016/j.compbiolchem.2015.09.009>
78. Speck-Planche A, Kleandrova VV, Ruso JM, DS Cordeiro MN (2016) *J Chem Inf Model* 56(3):588–598. <http://doi.org/10.1021/acs.jcim.5b00630>
79. Wiczorek M, Abualrous ET, Sticht J, Álvaro-Benito M, Stolzenberg S, Noé F, Freund C (2017) *Front Immunol* 8:292. <https://doi.org/10.3389/fimmu.2017.00292>
80. Toropova AP, Raškova M, Raška I Jr, Toropov AA (2021) *Theor Chem Acc* 140(2):1–8. <https://doi.org/10.1007/s00214-020-02707-8>
81. Liu P, Long W (2009) *Int J Mol Sci* 10(5):1978–1998. <https://doi.org/10.3390/ijms10051978>
82. Tong J, Li L, Bai M, Li K (2012) *Mol Inform* 36(5–6):1501023. <http://doi.org/10.1002/minf.201501023>
83. Masand VH, Rastija V, Patil MK, Gandhi A, Chapolikar A (2020) *SAR QSAR Environ* 31(9):643–654. <https://doi.org/10.1080/1062936X.2020.1784271>
84. Jenssen H, Fjell CD, Cherkasov A, Hancock RE (2008) *J Pept Sci* 14(1):110–114. <https://doi.org/10.1002/psc.908>
85. Mahmoodi-Reihani M, Abbasitabar F, Zare-Shahabadi V (2020) *ACS Omega* 5(11):5951–5958. <https://doi.org/10.1021/acsomega.9b04302>
86. Liu C, Zhang X, Nguyen TT, Liu J, Wu T, Lee E, Tu XM (2022) *Gen Psychiatr* 35(1). <http://doi.org/10.1136/gpsych-2021-100662>
87. Sukumar N, Prabhu G, Saha P (2014) *Applications of metaheuristics in process engineering*. Springer, Cham, pp 315–324
88. Andrade-Ochoa S, García-Machorro J, Bello M, Rodríguez-Valdez LM, Flores-Sandoval CA, Correa-Basurto J (2018) *J Biomol Struct Dyn* 36(9):2312–2330. <https://doi.org/10.1080/07391102.2017.1352538>
89. Zhou X, Li Z, Dai Z, Zou X (2010) *J Mol Graph Model* 29(2):188–196. <https://doi.org/10.1016/j.jmgm.2010.06.002>
90. Taboureau O (2010) *In antimicrobial peptides*. Humana Press, Totowa, pp 77–86. http://doi.org/10.1007/978-1-60761-594-1_6
91. He R, Ma H, Zhao W, Qu W, Zhao J, Luo L, Zhu W (2012) *Int J Pept*. <http://doi.org/10.1155/2012/620609>
92. Rajkumar M, Bhukya SN, Ahalya N, Elumalai G, Sivanandam K, Almutairi K, Alonazi WB, Soma SR, Urugo MM (2022) *BioMed Res Int*. <http://doi.org/10.1155/2022/7760734>
93. Zhou P, Liu Q, Wu T, Miao Q, Shang S, Wang H, Chen Z, Wang S, Wang H (2021) *J Chem Inf Model* 61(4):1718–1731. <https://doi.org/10.1021/acs.jcim.0c01370>

Part V
SMILES and Quasi-SMILES
for QSPR/QSAR

Chapter 12

SMILES and Quasi-SMILES Descriptors in QSAR/QSPR Modeling of Diverse Materials Properties in Safety and Environment Application



Yong Pan, Xin Zhang, and Juncheng Jiang

Abstract A brief summary of QSAR/QSPR methodology, together with an explanation of the approach using SMILES and quasi-SMILES descriptors to study diverse hazardous characteristics of diverse materials, is given. Studies of several properties of importance to safety and environment application are described including (i) the cytotoxicity of heterogeneous single metal oxide-based engineered nanoparticles, (ii) the cytotoxicity of a series of metal oxide nanoparticles, (iii) the flammability properties of chemicals and their mixture, (iv) thermal hazards properties of ionic liquids and their mixture and (v) the toxicity of ionic liquids and their mixtures. The limitations and outlook of this field in safety and environment are discussed.

Keywords QSAR/QSPR · SMILES · Toxicity · Nano-metal oxide · Flammability properties · Ionic liquids

12.1 Introduction

12.1.1 QSAR/QSPR Methods

Over the past few decades, cheminformatics has been emerging with the rise in information science and computational chemistry. Quantitative structure–property/activity relationship (QSPR/QSAR) is a hot research topic in cheminformatics. Combining the theoretical computational methods with various statistical tools, QSPR/QSAR is used to determine the physicochemical or biological properties as a quantitative function of the molecular structure. The basic assumption is that the physicochemical properties or activities are dependent on the molecular structure. This means that the properties or activities can be expressed as a function of the chemical structure. Taking the structure as the independent variable and the macroscopic

Y. Pan (✉) · X. Zhang · J. Jiang
College of Safety Science and Engineering, Nanjing Tech University, Nanjing, China
e-mail: yongpan@njtech.edu.cn

properties as the dependent variable, a quantitative relationship between them can be established by using mathematical and statistical methods. Based on the constructed QSPR/QSAR models, it can be used to predict various properties of new or unsynthesized compounds [1–3]. It can be also possible to identify the key structural factors of molecules that determine the macroscopic properties. Therefore, it is helpful to reveal the underlying mechanism and design the molecular structure to improve the property or activity.

QSPR/QSAR has been widely used to predict the biological activity and toxicity, the metabolic kinetic parameters of drugs, the physicochemical properties and the environmental effects [1, 2, 4–7]. This research covers many disciplines such as chemistry, medicine, life sciences and environmental sciences. QSPR/QSAR can significantly reduce research time and costs, which is of both theoretical and practical significance. Therefore, QSPR/QSAR has been increasingly applied to the design of chemical processes, the design of drug molecules and the evaluation of environmental risks.

12.1.2 Brief Description of the QSAR/QSPR Methodology

A typical QSAR/QSPR study contains the main steps as below.

- (1) Data collection: It includes various physicochemical properties and structural data from databases, manuals or experimental measurements.
- (2) Description of the molecular structure: According to certain theories or rules, structural parameters that reflect various structural information can be calculated, such as topological, compositional and quantum chemical parameters.
- (3) Selection of the molecular descriptors: The characteristic structure parameters should be closely related to the target properties, which are identified as molecular descriptors. Therefore, various statistical methods and optimization algorithms are applied to extract the characteristic molecular descriptors from a large number of structure parameters.
- (4) Construction of prediction model: The prediction models including regression methods, neural networks and support vector machines are often used to build a quantitative relationship between the selected molecular descriptors and the target properties.
- (5) Model evaluation and validation: The reliability of the constructed QSAR/QSPR model and the predictive capability of the model are evaluated by the mean correlation coefficient (R^2) and root mean square error (RMSE).

Among these steps, the description of molecular structure, the selection of molecular descriptors and the construction of prediction model are three key steps, which will be described as below.

12.1.2.1 Molecular Descriptors

The selected molecular descriptors play a key role in the quality of the model. Commonly used molecular descriptors can be divided into two main categories: experimental descriptors and theoretical descriptors. Early QSPR/QSAR studies often used a number of experimental descriptors, such as octanol–water partition coefficient, water solubility, Hammett’s constant and Taft’s constant. The advantage of these descriptors is that the physicochemical meaning is clear, and the disadvantage is that the acquisition of these parameters is labor intensive and costly.

With the development of knowledge in mathematics, molecular topology, quantum chemistry and other disciplines, theoretical molecular descriptors have been developed rapidly. Compared with experimental descriptors, theoretical descriptors have the following advantages: (1) Instead of the experimental characterizations, only the structural information of the molecule is required, which makes it possible to study the properties of unsynthesized compounds and greatly expands the application scope of QSPR; (2) the acquisition of these parameters is not restricted by experimental conditions, which is more convenient and faster. Moreover, the accuracy and speed of the calculations have also been improved with the development of computer technology; (3) these parameters provide a more comprehensive and detailed description of the molecular structure, which is beneficial to reveal the underlying mechanisms [8, 9].

12.1.2.2 Descriptor Selection Methods

In QSAR/QSPR studies, if the underlying mechanism is unknown, as many molecular descriptors as possible are often chosen to avoid omitting the significant factors. From the above-mentioned discussions, many types of molecular descriptors can be calculated. Such a large number of structural parameters must contain a large amount of useless and repetitive information for modeling, which affects and interferes with the construction and interpretation of QSPR models. In order to build QSPR models with fine fitting, predictivity, stability and interpretation, it is necessary to effectively identify and filter the molecular descriptors. The commonly used selection methods are listed as follows.

Multiple Linear Regressions-Based Selection Methods

The multiple linear regressions-based selection methods take the significance of the molecular descriptors on the model as a criterion. The criterion is that the addition or elimination of the descriptor has a significant effect on the model and the other molecular descriptors, which should also meet a predetermined significance level. There are three main types of such kind of methods including the forward selection, backward elimination and stepwise regression [10, 11].

The forward variable selection method starts with a one-parameter model and gradually increases the model parameters according to the significance criterion. The backward variable removal method starts with a model with all numerator descriptors and gradually decreases the model parameters according to the significance criterion. The stepwise regression method is a two-way selection regression that both adds and removes parameters from the model. It achieves two-way selection by setting two significance level criteria. One criterion is used to include the descriptor in the model, and the other criterion is used to remove the descriptor.

These methods are suitable for variable selection and model optimization for data where there is no multicollinearity between variables. The advantages are that they are simple and intuitive. The procedures are easy to implement, and the corresponding solutions can be obtained quickly. The disadvantage is that they cannot traverse all combinations of variables, which does not guarantee that the optimal solution in the variable space is found. When variable selection is performed on a large amount of data, these methods often result in a locally optimal solution.

Model Fitting-Based Selection Methods

This type of method often uses the goodness of fit of the model as a criterion for the simulation screening of variables. Such methods include optimal multiple linear regression and heuristic regression [1, 3].

The optimal multiple linear regression method first finds all orthogonal pairs in the initial set of descriptors. These orthogonal pairs are then used separately to model the physical properties of the target, resulting in a series of two-parameter models. The remaining descriptors that are not colinear with the parameters of several of the models with the largest degree of fitting are then added to the model one by one, resulting in a series of three-parameter models. If the degree of fitting of each of these three-parameter models is less than that of the two-parameter model with the largest degree of fitting, then the two-parameter model is the final result. Otherwise, the model variables continue to be added as described above until the optimal result is produced.

The heuristic regression method first calculates all the one-parameter models, removing the parts of them where the degree of fitting and significance is smaller than the set criteria. All two-parameter models are calculated from the retained numerator descriptors. The molecular descriptors with smaller parameter correlations to the part of the model with the largest degree of fitting are selected and added to the model resulting in a series of three-parameter models. The models with the largest degree of fitting were then selected. The model parameters are gradually increased as described above until the desired model size is reached and the model with the largest degree of fitting is selected as the final result.

Both methods are fast and unlimited in the size of the dataset and often result in a globally optimal solution. In comparison, the optimal multiple linear regression method is faster than the heuristic regression method.

Search Algorithms-Based Selection Methods

The main disadvantage of the above-mentioned methods is that they do not have global search capability and thus do not guarantee a globally optimal solution. In contrast, search algorithms such as simulated annealing algorithms and genetic algorithms (GAs) have considerable search capabilities. When they are combined with modeling methods such as multiple linear regression, partial least squares and artificial neural networks, they are able to search for the optimal model in the variable space within a limited time under certain conditions. Such methods have received great attention from researchers in recent years and have been better applied in QSPR research.

The simulated annealing algorithm is a relatively new optimization algorithm, which is derived from the solid annealing principle. The algorithm starts from the initial solution and the initial values of the control parameters, repeats the iterative process of “generate a new solution → calculate the objective function difference → accept or discard” for the current solution and gradually decays the values of the control parameters. It is a stochastic search algorithm based on the Monte Carlo iterative solution method, which has the potential to achieve global optimality and avoid local optimality. Therefore, it has been successfully used in QSPR studies of organic matter. For example, Jurs group [12] at Pennsylvania State University has combined simulated annealing algorithms with artificial neural networks for the selection of molecular structure parameters. They applied them to QSPR studies of many physical and chemical properties, achieving many interesting results.

Genetic algorithm (GA) is an adaptive global optimization probabilistic search method that simulates the genetic and evolutionary processes of organisms in their natural environment. It was first proposed by Holland in 1960 [13]. Based on the Darwin's fundamental principle of biological evolution in nature, superiority and inferiority produce individuals more adapted to their environment through crossover and mutation of genes. This principle is used to find the optimal answer to a practical problem and finally to obtain the optimal answer to a problem. GAs consist of three genetic operons: replication, hybridization and mutation. The evolutionary process is carried out by genetic operons. Genetic operators translate genetic concepts such as selection, recombination (or crossover) and variation into data processing to solve optimization problems dynamically. The problem is solved by so-called artificial chromosomes, which are changed and adapted by the optimization process until an optimization goal is obtained. The chromosomes contain information called genes, which are usually represented by strings. Depending on the problem to be solved, the string can be binary, an integer or even a real number.

GA is a simple, flexible, common and efficient global optimization algorithm. It performs parallel searches along multiple routes and generally does not fall into the trap of local optimality. It is able to find the global optimal solution among better local solutions. As a result, the study and application of GA have now become a dynamic direction internationally, with successful applications in process control, fault diagnosis, nonlinear fitting and many other engineering and research areas.

In 1994, Rogers and Hopfinger [14] introduced GAs into QSPR research for the first time. The GAs are used to intelligently select a reasonable combination of variables to obtain the optimal model. The main steps in QSPR research based on GAs are as follows

(1) Generating initial groups

First, an initial group set is generated. Once the initial groups are generated, each individual is evaluated using a score function.

(2) Selecting operation

A key feature of GAs is that only the optimal chromosomes pass on their characteristics to the next generation during evolution. Once all individuals in the group have been evaluated, the individuals to be retained in the new group can be selected based on the scores of the individuals in the group combined with a random method. For each individual to be eliminated, a new individual will be substituted. Commonly used selection functions are roulette selection, league selection and truncated selection. In roulette selection, the probability of selecting each individual is proportional to its score (fitness); in league selection, individuals are selected from the group to compete against each other, with the highest scoring individuals being retained; in the truncated selection, individuals are first ranked in order of their score and the optimal ones are selected.

(3) Crossbreeding operations

To perform the crossover operation, two retained individuals are selected as females in the group, then the two females are randomly divided into two segments, and a portion of the different females is later selected to form a new individual.

(4) Variation operations

The mutation operation, in which an individual is randomly selected in the group and an element of that individual is randomly changed to produce a new individual, results in a new property. All individuals generated by these two steps are evaluated using the score function, and new individuals are then selected according to their scores, resulting in a new group.

(5) Comparing operations

In order to preserve the optimal individuals, the optimal groups are used to preserve them. After the crosses and mutations have been made, the individuals of the new group are compared with those of the optimal group one by one, and if there are better individuals in the new group, they are copied into the optimal group.

(6) Convergence judgment

There are three ways to determine whether the calculation is converged: (1) The number of cycles is defined. When the number of steps has reached the defined value, the calculation is considered to be converged; (2) the total score of the optimal group is defined. When the total score of the optimal group no longer changes after a number of genetic operations, the calculation can be considered to be converged; (3) the average score of the group is defined. If

the average score of the group maintains constant for a number of times, the calculation is considered to be converged.

Compared to other methods, QSPR studies based on GA selection variables have three advantages: (1) the ability to find a set of models effectively, whereas other methods often provide only a single model. (2) The fitness function is not constrained by conditions such as continuity and differentiability and has a wide range of applicability. (3) It has inherent implicit parallelism and a good global search capability. (4) To build the model of multiple forms of linear combinations, the mathematical transformation of variables can be defined. In particular, these parameters can be classified by building truncated models, to obtain more useful information.

Because GA has a considerable search capability, when it is combined with modeling methods such as multiple linear regression, partial least squares and artificial neural networks, it is able to search for the optimal model in variable space in a limited time under certain conditions. Therefore, in recent years, GAs have received a great deal of attention and have been better applied in QSPR research [10].

It can be concluded that each of the above-mentioned variable selection methods has its own advantages, disadvantages and scope of application. Generally speaking, for problems with a linear relationship between the response variable and the independent variable, stepwise regression, heuristic regression and variable optimal subset regression are mostly used. However, for complex nonlinear problems, variable selection based on GAs often gives more satisfactory results.

12.1.2.3 Modeling Methods

To build quantitative functional relationships between the properties/activities and the molecular descriptors, the selected mathematical methods are a major step in QSAR/QSPR research. The commonly used modeling methods are divided into two main categories: linear methods such as multiple linear regression, principal component regression and partial least squares regression and nonlinear methods such as artificial neural networks and support vector machines.

Multiple Linear Regression Method

Multiple linear regression methods are the most common statistical method used in traditional QSAR/QSPR studies. The multiple linear regression process is the process of establishing a linear expression between the response variable and multiple independent variables. Assuming that there are m molecular descriptors, denoted by x_1, x_2, \dots, x_m , and the target materiality is denoted by y ; and there are n sample compounds, x_1, x_2, \dots, x_m, y are all n -dimensional vectors. Multiple linear regression refers to the establishment of a linear relationship between y and x_1, x_2, \dots, x_m as below.

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_mx_m \quad (12.1.1)$$

where b_0, b_1, \dots, b_m are the constants. b_0, b_1, \dots, b_m are obtained by solving a system of linear equations. The least squares parameter estimation is usually used to minimize the sum of squared errors. The molecular descriptors of the sample compounds form a matrix of coefficients that can only be solved when the matrix is full rank.

The multiple linear regression method is easy to use with the intuitive model, which is favorable to obtain the underlying mechanism. The disadvantage is that the resulting linear regression model may be distorted when the system is noisy or disturbed.

Principal Component Regression Method

The principal component regression is a linear combination of the original molecular descriptors to obtain principal components, which act as estimation parameters to build a multivariate linear model of their relationship with the target properties. Therefore, it is a combination of principal component analysis and multiple linear regression.

The purpose of principal component analysis, also known as factor analysis, is to obtain new variables of comparable variability but small dimensionality by linearly combining the original variables, which are known as principal components or factors. This process is achieved through matrix transformation. The principal components are inherently uncorrelated and can therefore be used directly in linear regression modeling.

The main steps in principal component regression include (1) standardization of the data, (2) derivation of the eigenvectors from the covariance matrix of the data, and (3) selection of principal components for multiple regression analysis.

The advantage of principal component regression is that it can effectively solve the problem of multicollinearity among variables by combining and filtering the information in the original data; the disadvantage is that it only deals with the independent variables and does not consider the information of the response variables, so the first principal component it obtains does not necessarily have the strongest correlation with the response variables. For this reason, it has been improved by introducing the partial least squares regression method.

Partial Least Squares Method

The partial least squares method is also a regression method based on component extraction [15]. Unlike principal component regression, it combines the extraction of principal components with the target properties to ensure that the principal components are correlated with the target properties. The process involves extracting components from both the independent variable data and the respondent data, which should

meet two requirements: (1) The extracted components represent as much information as possible from the original data table; (2) the correlation between the extracted components from the independent variable data and the respondent data is maximized, and the extracted components are then used to model the regression. If the model meets the modeling requirements, the component extraction operation is terminated; otherwise, the components are extracted again from the remaining data information, and these components must also meet the two requirements above. Then, the extracted principal components are modeled again. This process is repeated several times until the modeling requirements are met. The model is then reduced to a model of the original variables.

Compared with the traditional multiple linear regression and principal component regression methods, the partial least squares method has the following advantages: (1) The original data information is integrated and filtered, effectively solving the problem of multicollinearity among variables; (2) when the number of independent variables is more than the number of samples, statistically significant equations can still be obtained; (3) both the information of the independent variables and the response variables are considered, making it easier to obtain meaningful; (4) the use of interaction tests to select the optimal number of principal components in the model reduces the "chance correlation" of the model. Because of these obvious advantages, the partial least squares method has good robustness and strong predictive power. The partial least squares method has become one of the more commonly used modeling methods in QSAR/QSPR studies of organic matter.

Artificial Neural Network Method

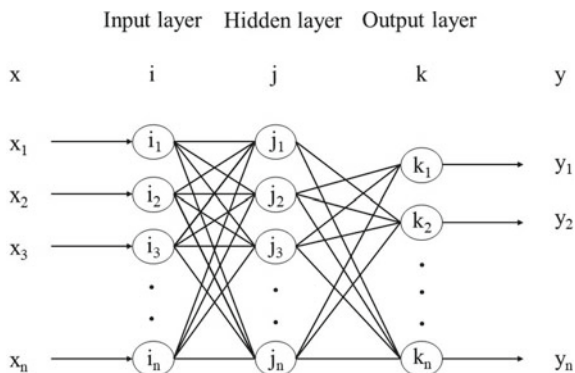
The artificial neural network is a nonlinear, adaptive information processing system composed of a large number of interconnected processing units. It is proposed on the basis of modern neuroscience research results and attempts to process information by simulating the way the brain's neural network processes and remembers information.

According to the different learning strategies, artificial neural networks can be divided into two categories: supervised neural networks and unsupervised neural networks. The supervised neural networks are mainly trained on known samples and then predict the unknown samples. Unsupervised methods, also known as self-organizing neural networks, can be used to classify compounds without training on known samples, such as Kohonen neural networks and Hopfield models. Currently, BP neural networks are the most used in QSAR/QSPR research.

BP neural networks generally adopt a three-layer network structure, i.e., input layer, implicit layer and output layer. The input layer receives the external data input, the implicit layer processes and transforms the input data, and the output layer produces the output results. A typical BP network structure model is shown in Fig. 12.1.

Each layer of the network contains a number of neurons, with the number of neurons in the input and output layers determined by the number of variables in the model and the number of neurons in the hidden layer determined by trial and error.

Fig. 12.1 Structure model of BP neural network



Each neuron in the implicit and output layers contains two functions: a summation function and a transfer function. The sum function is a weighted sum of all input neurons entering each hidden layer neuron and converts the result into a single value for further processing in the transfer function; the transfer function is used to convert the summed information into output. The sigmoid transfer function is most widely used as below.

$$F(x) = \frac{1}{[1 + \exp(-x)]} \quad (12.1.2)$$

The specific steps of the BP algorithm are briefly described as follows: (1) initialization. The coefficients and values of the weights of each layer are randomly set; (2) the training sample data X is added to the input layer of the network, and the output Y of each layer is calculated. The error is obtained by comparing the output with the expected value; (3) the connection weights according to the error are readjusted; (4) if it is less than the predetermined error, the network is considered to be converged and stops learning. Otherwise, it returns to Step (2) and continues to Step (3).

Artificial neural networks have many advantages such as nonlinearity, self-learning, adaptability, fault tolerance, associative memory and trainability, which are superior to traditional multiple linear regression and partial least squares and have become an important algorithm in QSAR/QSPR research. However, in the process of practical application, the neural network method also reveals the following shortcomings: (1) Due to the strong nonlinear fitting ability of neural networks, when the training set samples are small, the phenomenon of “overfitting” often occurs; (2) the neural network is built as a “black box” model and the input and output are not the same. The relationship between input and output is unclear; (3) due to the randomness of the initialization of the neural network weights, the results are difficult to repeat.

The existence of these problems limits the further application of neural networks in QSAR/QSPR research, and new and more superior machine learning algorithms need to be introduced to promote the profound development of QSAR/QSPR research.

Support Vector Machine Method

(1) Theoretical background

The support vector machine (SVM) algorithm is a new machine learning method proposed by Vapnik and his co-workers [16, 17] in 1995, based on statistical learning theory.

The term “statistical learning theory” refers to a theory that specializes in the study of machine learning patterns in the context of small samples. Vapnik et al. [16] started to work on this area in the 1960s and 1970s, and by the mid-1990s, as their theory continued to develop and mature, the theory began to gain increasing attention. The traditional statistical approach regards empirical risk minimization (ERM) as the starting point, without examining theoretical issues such as its rationality, applicability and achievable quality of approximation. It finally makes the empirical risk minimization not guarantee expected risk minimization. Unlike the statistical learning theory, it proposes the principle of structural risk minimization and the core concept of VC dimension. It also states that to minimize the expected risk, both the empirical risk and the VC dimension must be minimized. VC dimension theory provides a rigorous justification for the ERM principle, i.e., a sufficient condition for consistent convergence, a sufficient condition for fast convergence and a sufficient condition for consistent convergence independent of the probability distribution. Therefore, it has a rigorous theoretical foundation. It is on this theoretical basis that the SVM approach is developed. To obtain the optimal universality, it is based on VC dimensional theory and the principle of structural risk minimization and seeks the optimal compromise between the complexity of the model (i.e., the learning accuracy for a given training sample) and the learning ability (i.e., the ability to identify arbitrary samples without error) based on limited sample information.

Compared with traditional statistical learning methods, the SVM method has the following main advantages [18]. (1) It has a strict theoretical and mathematical foundation, overcoming the “empirical” nature of traditional methods; (2) it is specifically designed for the finite sample case and its optimal solution is based on the information of the available samples, rather than the optimal solution when the number of samples tends to infinity; (3) the algorithm is ultimately transformed into a convex optimization problem, so the solution of SVM is globally unique, solving the local minimum problem that cannot be avoided by neural networks; (4) by applying the kernel function technique, the nonlinear problem in the input space is mapped to the high-dimensional feature space through the nonlinear and linear function in the high-dimensional feature space which is constructed to realize the nonlinear function in the original space. Therefore, the model has a good universality. The complexity of the algorithm is closely related to the dimensionality of the input vector, thus avoiding the “dimensionality disaster”.

Therefore, SVM has become an international research hotspot. In an article published in Science, SVMs are “a very popular approach and success story in the field of machine learning and a very compelling direction for development”.

(2) Mechanism

SVM algorithms were originally applied to solve classification problems. In recent years, with the introduction of the ϵ -insensitive loss function, SVM algorithms have also been increasingly used to solve regression problems and have shown good performance. In this paper, we focus on the application of SVM to regression problems, so the following is a brief introduction to the SVM regression algorithm and we do not go into the classification methods. The detailed principles of both can be found in the SVM user guidance.

The core idea of the SVM regression algorithm is to find an optimal hyperplane that minimizes the distance from all sample points to the hyperplane, as illustrated in Fig. 12.2. As can be seen from Fig. 12.2, the optimal hyperplane is actually determined by a small number of samples called support vectors.

We assume that the training sample set $\{(x_i, y_i), i = 1, \dots, n\}$ is given, where $x_i \in R_n$ is the input value of the i th learning sample and $y_i \in R$ is the corresponding target value. For linear regression, a linear function is applied for estimation.

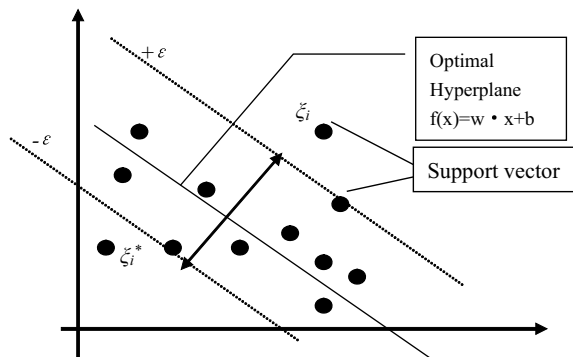
$$f(x) = (w \cdot x) + b \tag{12.1.3}$$

To ensure that Eq. (12.1.3) is flat, a minimum w must be found. Assuming that all training data (x_i, y_i) can be fitted with a linear function at accuracy ϵ , the problem of finding the minimum w is transformed into minimizing the model complexity, which is shown below:

$$\min \frac{1}{2} \|w\|^2 (y_i - w \cdot x - b \leq \epsilon, w \cdot x + b - y_i \leq \epsilon) \tag{12.1.4}$$

Taking the fitting error into account, a relaxation factor $\xi \geq 0, \xi^* \geq 0$ and a penalty factor C are introduced and the corresponding quadratic programming problem is

Fig. 12.2 SVM for regression



rewritten as

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*)$$

$$(y_i - w \cdot x - b \leq \varepsilon + \xi_i, w \cdot x + b - y_i \leq \varepsilon + \xi_i^*, \xi_i, \xi_i^* \geq 0) \quad (12.1.5)$$

The penalty factor $C > 0$ is used to balance the flatness of the regression function $f(x)$ and the number of sample points with deviations greater than ε . Equation (12.1.5) is derived based on the following ε -insensitive loss function. $|\xi|_\varepsilon$ is expressed as follows

$$|\xi|_\varepsilon = \begin{cases} 0 & (|\xi| \leq \varepsilon) \\ |\xi| - \varepsilon & (\text{otherwise}) \end{cases} \quad (12.1.6)$$

When the number of samples is small, the above SVM is generally solved using pairwise theory, which transforms it into a quadratic programming problem. The following Lagrange equation is developed:

$$l(w, \xi, \xi^*) = \frac{1}{2}(w \cdot w) + C \sum_{i=1}^n (\xi_i + \xi_i^*) - \sum_{i=1}^n \alpha_i (\varepsilon + \xi_i$$

$$+ y_i - \langle w, x_i \rangle - b) - \sum_{i=1}^n \alpha_i (\varepsilon + \xi_i^* + y_i$$

$$- \langle w, x_i \rangle - b) - \sum_{i=1}^n (\eta_i \xi_i + \eta_i^* \xi_i^*) \quad (12.1.7)$$

The partial derivatives of the above equation are equal to 0 for the parameters w , b , ξ_i , ξ_i^* , and the pairwise optimization problem is obtained by substituting Eq. (12.1.7)

$$\min \frac{1}{2} \sum_{i,j=1}^n (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle + \sum_{i=1}^n \alpha_i (\varepsilon - y_i) + \sum_{i=1}^n \alpha_i^* (\varepsilon + y_i)$$

$$\left(\sum_{i=1}^n (\alpha_i - \alpha_i^*) = 0, \alpha_i, \alpha_i^* \in [0, C] \right) \quad (12.1.8)$$

For nonlinear regression, the SVM solution is to map the sample into a high-dimensional feature space by a nonlinear mapping φ and solve it by conventional linear methods. Assuming that the sample X is mapped to a high-dimensional space using a nonlinear function $\varphi(X)$, the nonlinear regression problem is transformed into Eq. (12.1.9).

$$\min \frac{1}{2} \sum_{i,j=1}^n (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) (\phi(x_i), \phi(x_j)) + \sum_{i=1}^n \alpha_i (\varepsilon - y_i) + \sum_{i=1}^n \alpha_i^* (\varepsilon + y_i)$$

$$\left(\sum_{i=1}^n (\alpha_i - \alpha_i^*) = 0, \alpha_i, \alpha_i^* \in [0, C] \right) \quad (12.1.9)$$

and thus obtain $w = \sum_{i=1}^n (\alpha_i - \alpha_i^*) \phi(x_i)$.

A SVM can map samples to a high-dimensional feature space through a kernel function transformation, with the kernel function $K(x, x')$ satisfying $K(x, x') = \langle \phi(x), \phi(x') \rangle$. Thus, Eq. (12.1.8) is rewritten as

$$\min \frac{1}{2} \sum_{i,j=1}^n (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(x_i, x_j) + \sum_{i=1}^n \alpha_i (\varepsilon - y_i) + \sum_{i=1}^n \alpha_i^* (\varepsilon + y_i)$$

$$(12.1.10)$$

The introduction of kernel functions allows the function to be solved directly in the input space, bypassing the feature space, thus avoiding the need to compute nonlinear mappings ϕ . The four main types of kernel functions commonly used in SVMs today are linear kernels, polynomial kernels, radial basis kernels and sigmoid kernels.

(3) Parameter optimization

In order to obtain the optimal universality, the SVM needs to adjust the corresponding combination of parameters in the modeling process, i.e., choosing the appropriate kernel function, determining the parameters of the kernel function, the penalty factor C and the size of ε in the ε -insensitive loss function. The kernel function determines the distribution of the input vectors in the high-dimensional space and the optimal hyperplane to be found and therefore determines the predictive power of the SVM to a large extent. There is no unified method for the selection of the kernel function, which is basically determined by empirical methods. The most commonly used kernel function in practice is the radial basis form of the radial basis function (RBF) kernel function, which has a high learning efficiency and learning rate. For the RBF kernel function, the most important parameter is the width of the kernel function γ , which determines the amplitude of the kernel function and therefore to some extent the universality of the SVM. The penalty factor (C) is also another important parameter controlling the prediction performance of the SVM, which controls the balance between maximizing the bound and minimizing the training error. If the parameter is too small, underfitting of the training data will occur; if the parameter is too large, the training data will be overfitted. Therefore, C also affects the training speed and universality of the SVM. The optimal value of ε depends on the noise of the data, which is usually unknown, while the number of support vectors has to be considered in practical problems even if sufficient knowledge is available to choose the optimal value of ε . The ε -insensitive loss function prevents the entire training set

from reaching the boundary condition, allowing sparsity in the solution of the dyadic form, and therefore theoretically, it is also important to choose the right value of ε .

At present, there is no unified method to determine the optimal parameters of SVM. The commonly used methods are the single-factor rotation method and the grid point search (GS) method. The single-factor rotation method is based on a certain empirical basis to optimize the parameters under study one by one to find the optimal value, its advantage is that it can quickly build a model, but the disadvantage is that it does not consider the interaction between the parameters. The advantage of this method is that it is fast to build a model, but the disadvantage is that it does not take the interactions between the parameters into account. The grid point search method generally uses cross-validation to select the appropriate parameters through multiple trials.

12.2 SMILES and Quasi-SMILES Descriptors

Simplified molecular input line entry system (SMILES) is a specification for explicitly describing molecular structures using ASCII strings. SMILES was developed by Weininger and Weininger [19] in the late 1980s and has been modified and extended by others, notably Daylight Chemical Information Systems Inc.

The SMILES formula consists of a series of characters without spaces, and it is essential to ensure that the chemical structure of a substance corresponds to its SMILES expression. One substance corresponds to only one SMILES structure. Therefore, in the calculation of SMILES expressions for substances, certain grammatical expression rules are set for atoms, chemical bonds (single, double and triple bonds), branched chains, rings, atomic chirality, isotopes, etc. The specific expression rules are listed in Table 12.1.

When using the SMILES formula to represent the chemical structure of a substance, the hydrogen atoms in the chemical structure are first eliminated. If the chemical structure contains rings, the rings also need to be opened and represented by breaking them off. The atoms in the rings are all represented in lowercase letters. The two atoms connected at the ring break are marked with the same number to indicate that there is a bond between the atoms. The branched chains in the chemical structure are written in parentheses.

SMILES rules have recently become an international standard and are considered to be the most applicable and compatible form of linear coding compared to other rules. This is because SMILES can be used quickly to express the structural information of a compound into a computer-readable code, requiring only the atomic symbols of the compound, the bond symbols, and certain syntactic expression rules.

SMILES is calculated by a longitudinal priority traversal tree algorithm, which converts the chemical formula of a compound into a SMILES expression by means of a sequence of characters without spaces. The basic rules to be followed in the SMILES coding transformation are (i) the hydrogen atom is ignored during the chemical structure transformation; (ii) the aromatic ring structure is opened before

Table 12.1 Expression rules of SMILES

Type	SMILES expressions	Notes	Example
Atom	① [Element symbol]	Atoms such as C, N, O, P, S, Br, Cl and I of organic chemicals are omitted in the square brackets	Iron atom: [Fe]
	② Hydrogen atom is omitted		Water: O
	③ [Element symbol ± electric charge]	“+” and “-” denote positive and negative charges, respectively, followed by the charge value	The tetravalent titanium ion: [Ti+4]
Chemical bond	① Double bond is represented as “=”		Carbon dioxide: O=C=O
	② Triple bond is represented as “#”		Hydrogen cyanide: C#N
	③ The ring needs to be broken, and the two atoms at the break are marked with the same number	The C, O, S and N atoms in the aromatic ring are represented as lowercase letters	Cyclohexane: C1CCCCC1 Benzene: c1ccccc1
Branched chain	① Branched chain on the carbon chain is represented as “()”		Propionic acid: CCC(=O)O
Stereochemistry	① The structure on each side of the double bond is represented as “/” and “\”	“/” and “\” represent cis; “/” and “/” represent trans	Trans difluoroethylene: F/C=C/F Cis difluoroethylene: F/C=C\F
	② Chiral carbon atom is marked with “@” or “@@”		L-alanine: N[C@@H](C)C(=O)O
Isotope	① Isotopes are shown with the mass number written in front of the element symbol		Chloroform-d: [2H]C(Cl)(Cl)Cl

coding, or expressed in Kekuler style; (iii) in the opened chain expression, the number is used to mark the broken atom, the atom is represented by a lowercase letter, and the branched chain is characterized by round brackets. The SMILES strings are often used as input files in some calculation software and are converted into 2D or 3D structures, so that each compound has its own SMILES string structure. In addition, SMILES strings are compatible with a wide range of software and have been successfully applied to the toxicity prediction of traditional compounds.

SMILES is a traditional tool for representing the molecular structure. In contrast to conventional SMILES, quasi-SMILES can be used as a tool to establish quantitative features-property/activity relationships (QFPRs/QFARs) for endpoints that are defined not by molecular structure alone, but by a set of physicochemical and/or biochemical conditions.

The nano-QSAR study in this paper is divided into the following major steps.

- (1) Collection of sample data: The current data on physicochemical parameters (descriptors) and toxic effects of nanomaterials are mainly obtained from biological experiments, literature reports and authoritative databases.
- (2) Identification and acquisition of descriptors for metal oxide nanomaterials: The information on molecular structure, elemental periodicity and quantum chemistry of nanoparticles was studied to establish descriptors of physical and chemical parameters such as absolute molecular weight, particle size distribution, surface area, morphological parameters and zeta potential to characterize the physical and chemical characteristics of nanomaterials and to select descriptors that are closely related to the cellular toxicity of nanoparticles. The SMILES descriptors were combined with the SMILES structures of the nanomaterials to optimize and improve the SMILES descriptors and the optimized SMILES descriptors were used to characterize the basic structural information of the particles.
- (3) Screening of metal oxide nanomaterial descriptors: Firstly, the descriptors with high similarity were removed by correlation analysis to complete the pre-screening of descriptors. Then, support vector machine-recursive feature elimination (SVM-RFE) was jointly used to derive the importance ranking of subsets of descriptors. The optimal subset of features was determined according to the accuracy of the classification model.
- (4) Study and modeling of cytotoxicity of nanomaterials: Using the selected nanoparticle descriptors as input parameters and combining different modeling methods, nano-QSAR studies were conducted on the cytotoxicity of different nano-metal oxide systems to establish the corresponding toxicity classification and prediction models.
- (5) Evaluation and validation of the model: To evaluate and assess the fitting ability, stability and prediction ability of the model.
- (6) Mechanistic interpretation of the model: The model will be mechanistically interpreted to reveal the main factors affecting the cytotoxicity of different nanomaterials and their influence laws, to reveal the mechanism of toxicity of nanoparticles and to provide guidance for the synthesis and design of new nanomaterials. In summary, the characterization of nanomaterial structures, the calculation and screening of descriptors and the establishment of predictive models are the main contents of this study.

12.3 Study of Several Important Properties/Activities in Safety and Environmental Applications

12.3.1 *The Cytotoxicity of Metal Oxide Nanoparticles*

Nanotechnology is a symbol of science and technology in the twenty-first century. With the rapid development of nanotechnology, there are increasing types of nanomaterials and widespread applications. Nanomaterials, in a broad sense, are materials that have at least one dimension in the nanoscale range (1–100 nm, $1 \text{ nm} = 10^{-9} \text{ m}$) in three-dimensional space or are made up of the basic structural units of substances in this scale range. Nanomaterials are very small in size and have a very special structure and have many physical and chemical properties that are very different from those of macroscopic materials, such as large specific surface, very high reactivity and the unique surface effect, small size effect and macroscopic quantum tunneling effect of nanomaterials. With the industrialization of nanotechnology, nanomaterials are increasingly used in traditional and emerging industries such as the pharmaceutical industry, dyestuffs, coatings, food, cosmetics and environmental pollution control. However, this technology is a “double-edged sword”. While it brings great economic benefits and technological innovations, the safety issues arising from nanomaterials cannot be ignored, especially their biological toxicity, which has received widespread attention from researchers in various countries. There is a growing awareness of the enormous impact that atmospheric nanoparticles have on the environment and on living organisms. In addition to the atmospheric environment, nanoparticles are also present in local working environments, such as coal mining, welding and powder processing, where a large number of nanoparticles are floating in the surrounding environment and their impact on human health cannot be ignored. In addition, as nanomaterials are widely used in daily life, the possibility of contacting with nanomaterials for people has greatly increased. Either directly into the human body during production and use, or through the environment or food chain, nanomaterials have an inevitably negative impact on human health after an intrusion. It is found that many serious diseases can be caused by exposure to nanomaterials [20]. The toxicity of nanomaterials has become a major obstacle to the development of the nanotechnology industry. Therefore, the study of the biotoxicity of nanomaterials is an important issue that needs to be addressed in the development of nanotechnology and its industry.

In April 2003, Service [21] first published an article in Science on the biotoxic effects of nanomaterials. In the following year, researchers from various countries discussed the biotoxicity of nanomaterials and the potential environmental safety issues [22–24]. As a result, policies and measures have been taken to increase research on the biotoxicity of nanomaterials.

The determination of the cytotoxicity and safety of metal oxide nanomaterials has traditionally been carried out by experimental tests, and it is undoubtedly still the most effective way. However, traditional assays are controversial in terms of cost, efficiency and ethical implications and are not able to cope with the increasing number of newly developed nanomaterials on the market. With the development of

nanotechnology, many experimental data on the cytotoxicity of nano-metal oxides have emerged in recent years, but the difference in experimental conditions and methods between studies often makes it difficult to assess the toxicity of metal oxide nanoparticles. Furthermore, even though there are many toxicological methods available for assessing nanotoxicity, the effects of nanomaterials on cellular metabolism in vitro and in vivo are still unknown. Moreover, the inconsistent results between the various studies make it hard to develop a comprehensive system for studying the mechanisms of cytotoxicity.

The QSAR method is a simple and effective way to accurately predict the biological activity of a compound before it is synthesized. By converting the structural information of a compound into a descriptor and using mathematical calculations, the link between the descriptor and the target property is established, which is helpful to predict the relevant toxic effects and elucidate the mechanisms. Nano-QSAR is an extension of the traditional QSAR research and is a method to predict the bioactive effects of nanomaterials, which is a theoretical basis for the synthesis of new nanoparticles and the design of functional nanoparticles.

In recent years, optimized descriptors based on SMILES structures have also received a lot of attention from nano-QSAR researchers. With the emergence and development of CORAL software, Toropov et al. [25–28] proposed a series of conformational models for the study of the biotoxicity of nanomaterials, which facilitated the development of nano-QSAR research.

12.3.1.1 The Cytotoxicity of Single Metal Oxide Nanoparticles

Toropova et al. [29] established a model of malondialdehyde (MDA) levels in different organ wet tissues of rats under different effects of Al_2O_3 nanoparticles based on quasi-SMILES. The levels of MDA in different organ wet tissues were used as a standard measure of toxic effects. Numerical data on MDA concentrations in rat liver, kidney, brain and heart wet tissues were studied as endpoints, which were influenced by different doses, exposure times (3 and 14 days) and single oral treatments with 30 nm or 40 nm Al_2O_3 .

Manganelli et al. [30] developed a model to predict the survival of human embryonic kidney cells (HEK293) under 40 different experimental conditions using silica nanomaterials. They used SMILES-based descriptors as input parameters to the model and combined particle size, concentration and exposure time into the SMILES structure to form “quasi-SMILES”, thus fully characterizing the experimental conditions of the nanomaterials. The sample set was randomly divided into five groups, and then Monte Carlo optimization and modeling were carried out using CORAL software. The prediction models all had complex R^2 above 0.7, and the prediction results of the models were good.

Toropov and Toropova [31] developed a model based on quasi-SMILES to estimate the toxicity of ZnO nanoparticles to rats by intraperitoneal injection. They calculated the correlation weights of the quasi-SMILES fragments by the Monte Carlo method. A univariate toxicity model was developed with the numerical data of

the correlation weights. All available data were randomly divided into five parts, and the results of 36 experiments were divided into a training subsystem and a validation subsystem. The mean coefficient of determination was 0.957 (with a dispersion of 0.010 mg/kg), and the average root mean square error was 7.25 mg/kg (with a dispersion of 0.59 mg/kg). The method described is suitable for predicting the outcome of intraperitoneal injections of nanoparticles in rats and can also be used in other experiments which can be represented by quasi-SMILES, similar to the experiments described here.

12.3.1.2 The Cytotoxicity of a Series of Metal Oxide Nanoparticles

Toropova et al. [32] investigated the QSAR of the pLC50 for the toxic effects of 18 nano-metal oxides on *Escherichia coli* and used a Monte Carlo algorithm to develop a predictive model. The SMILES-based descriptor was obtained by combining the SMILES string calculated by ACD/ChemSketch software with the symbol “^” characterizing whether the cytotoxicity was photoinduced and was applied for the first time to the nano-QSAR model of nano-metal oxide cytotoxicity. The data were then randomly divided into training, calibration and validation sets with different functions according to a certain ratio. The stability of the constructed prediction models was verified.

Toropova et al. [33] developed a predictive model for cell membrane damage caused by a range of nano-metal oxides. They applied the optimal descriptors that were calculated from the so-called correlation weights for different concentrations and different exposure times. The numerical data of the correlation weights were calculated by Monte Carlo method. The results obtained are in good agreement with the experimental data. For the seven metal oxide nanoparticles, the chemical composition had the most important effect on cell membrane damage. Surprisingly, the effect of the dose on cell membrane damage was the lowest. Exposure time had a moderate effect on endpoints.

Pan et al. [34] coded some physicochemical properties related to the toxicity of nanomaterials into codes and formed a new string with the traditional SMILES structure. They proposed a new descriptor, namely the improved SMILES-based descriptor, which can characterize the structure of nanomaterials more comprehensively and easily. In this study, two nano-QSAR prediction models were developed for different nano-metal oxides targeting the toxicity effects of human keratinocytes and *Escherichia coli*, respectively. The average R^2 of the two models was as high as 0.95, and the models were rigorously validated for stability, predictive power and robustness. The mechanistic interpretation of the models was that the original particle size and hydrated particle size were the main factors for the biotoxicity of the nanomaterials.

Toropova et al. [29] developed a single QSAR model for predicting the cytotoxicity of metal oxide nanoparticles against (i) *Escherichia coli* (*E. coli*) and (ii) human keratinocyte cell lines (HaCaT) based on data on the half-lethal concentrations of 32 metal oxides nanoparticles. The mean R^2 and root mean square error (RMSE)

for the training set were 0.79 and 0.216; the R^2 and RMSE for the validation set were 0.90 and 0.247, respectively. The method yielded reasonably good models for compromised data related to the cytotoxicity of metal nanoparticles against *E. coli* and HaCaT.

Choi et al. [35] collected a large amount of toxicity data from the S2NANO (www.s2nano.org) database and developed a QSAR model for predicting the cell viability of 21 metal oxide nanomaterials on human lung bronchial epithelial cells and human dermal keratinocytes. The physicochemical properties of the nanomaterials and experimental conditions were transformed into codes, which combine the SMILES structures to form the quasi-SMILES descriptors. The effects of different coding methods on the performance of the nano-QSAR model were compared. It was shown that the QSAR models generated using the hierarchical clustering analysis (HCA) method had better performance than the min–max method.

Cao et al. [36] examined the LC50 of 21 nano-metal oxides on A549 cells by biological screening experiments to determine the nanotoxicity characteristics of the nanoparticles. A corresponding quantitative structure–activity relationship model for nanoparticles (nano-QSAR) was developed for the risk assessment of nano-metal oxides using an improved SMILES-based optimal descriptor and MC-PLS modeling approach. In addition, the effects and mechanisms of different physicochemical properties on their acute cytotoxicity are discussed. The R^2 and Q^2_{LOO} values of all four models were above 0.8, while all external validation coefficients of Q^2_{Ext} were above 0.7, indicating that all four models were reliable, stable and had satisfactory predictive power. The applicability and reliability of the improved SMILES-based optimal descriptors in predicting the acute cytotoxicity of the novel nano-metal oxides were also verified. Furthermore, the effects of structural factors on the acute cytotoxicity of nano-metal oxides showed that individual size and aggregation size were the most critical physical factors affecting the acute cytotoxicity of nano-metal oxides to A549 cells, followed by cat ion charge and zeta potential, with weaker effects of metal mass fraction and molecular weight. ROS experiments in A549 cells showed that the reactive oxygen species theory (mechanism I) in nano-metal oxides predominated in the mechanism of toxicity to A549 cells. In addition, the developed model has potential applications in guiding risk assessment and safer and greener design of nanomaterials and can be prioritized in virtual screening. The study of acute cytotoxicity of nano-metal oxides on A549 cells will also contribute to medical development.

Ahmadi [37] researched and developed a nano-QFAR (quantitative nano-featured activity relationship) model to predict the cell viability of metal oxide nanoparticles (MO-NPs) by applying quasi-SMILES such as cell line, assay method, exposure time, concentration, nanoparticle size and metal oxide type. A total of 83 quasi-SMILES of metal oxide nanoparticles were randomly divided into three sets: training set, validation set and test set. The results of the statistical models based on the equilibrium-related target function (TF_1), the exponential-desirability-related target function (TF_2) and Monte Carlo optimization were compared. The comparison of the results of the two objective functions showed that TF_2 improved the predictability of the model. The significance of the various trade-off features for increases and decreases in cell survival is provided. A mechanistic explanation of the important

factors of the model is also presented. The full statistical quality of the three TF_2 -based nano-QFAR models suggests that the developed models can be used to predict the cell viability of MO-NPs.

Toropova et al. [38] analyzed the sustainable nanotechnology (S2NANO) dataset containing 574 experimental cell viability and toxicity data points measured under different conditions for Al_2O_3 , CuO , Fe_2O_3 , Fe_3O_4 , SiO_2 , TiO_2 and ZnO . They used the quasi-SMILES molecular representation to develop a QSAR model based on classification and regression. The introduced quasi-SMILES takes all available information into account, including the structural characteristics of the nanoparticles (molecular structure, core size, etc.) and relevant experimental parameters (cell line, dose, exposure time, assay method, hydrodynamic size, surface charge, etc.). The resulting regression models showed adequate predictive power, while the classification models showed higher accuracy. As the analyzed datasets reported cell viability and cytotoxicity measured under a variety of experimental conditions, the developed models were able to capture the general safety profile of the seven types of nanoparticles.

The antibacterial activity and cytotoxicity of metal oxide nanoparticles are known to be determined by the energy band gap. Toropova and Toropov [39] gave prediction models for the energy gap (E_g) based on quasi-SMILES nano-QSARs for E_g of metal oxide nanoparticles. The new version of quasi-SMILES has been applied to model the energy band gap of metal oxide nanoparticles. Both the correlation index and the correlation strength index have the potential to improve the prediction potential of nano-QSAR for the energy band gap of metal oxide nanoparticles. However, calculations using three different data show that the correlation intensity index gives a more reliable model for the prediction of the energy band gap of metal oxide nanoparticles.

12.3.2 Flammability Properties of Chemicals and Their Mixtures

Although the flammability properties contain the flash point (FP), auto-ignition temperature, and flammability limits, etc., the current QSPR research with SMILES and quasi-SMILES descriptors only focused on the FP. Saldana et al. [40] developed a QSPR model using the SMILES molecular representation to model the FP and cetane number (CN) of molecules that may be found in alternative fuels. The models are applicable to hydrocarbons, alcohols and esters. A database containing FP and CN for these types of molecules has been created using experimental data from the available literature. For both properties, various methods of linear modeling approaches including GAs and PLS and nonlinear approaches including feed-forward artificial neural networks (FF-ANNs), generalized regression neural networks (GRNNs), SVMs and graph machines (GMs) have been investigated. For both properties, none of the models obtained was more accurate than the others. Therefore, the consensus

modeling was proposed, which improves robustness and predictability compared to individual models. The results were that FP depends mainly on the total number of carbon atoms in the molecule. They also show how the CN evolves when one or two alcohol groups are added to a carbon chain and when these are moved along the chain.

Toropova et al. [41, 42] applied quasi-SMILES to model the flammability of binary and ternary liquid mixtures separately. The method provides a good model for predicting the flash points (in degrees Celsius) of binary and ternary mixtures of organic substances. The associated ideality index (IIC) is a criterion for the predictive potential of the QSPR/QSAR model. The application of the IIC to improve the flammability model for ternary liquid mixtures confirms the applicability of this criterion to improve the predictive potential of the above models.

Gantzer et al. [43] compared their work with that of Saldana et al. [40]. In the work of Saldana et al., the database was filtered to retain only compounds of interest, such as hydrocarbons and oxygenated molecules (mainly alcohols and esters). In the work of Gantzer et al., they considered the complete database including additional families of compounds such as aldehydes, ketones, ethers and alkynes. This database of 785 chemicals was randomly divided into two subsets, 599 compounds for training and 186 for testing the model. They calculated ISIDA descriptors to encode molecular features based on SMILES. For each descriptor set, the parameters of the support vector regression (SVR) were optimized using fivefold cross-validation (5-CV). The models based on two to four atomic sequences and their built descriptors performed well according to internal (cross-validation) and external validation. The model of Gantzer et al. showed a similar performance to that derived by Saldana et al. The small difference in performance can be attributed to the Gantzer et al. database, which contains a wider diversity than the database used by Saldana et al. and the use of a single QSPR, whereas Saldana et al. used several QSPRs in a consensus model.

12.3.3 *Thermal Hazard Properties of Ionic Liquids and Their Mixtures*

Thermal hazards have become one of the fundamental characteristics of different ionic liquids (ILs). The thermal decomposition of ionic liquids (ILs) is also an important aspect in the evaluation of the thermal hazards of ILs.

Lotfi et al. [44] focused on predicting the thermal decomposition (T_d) of ionic liquids (ILs). They developed QSPR models for the molecular structure of ILs based on the SMILES notation and used the Monte Carlo algorithm of the CORAL software to calculate T_d for 263 imidazole-like ionic liquids. They constructed four QSPR models with a hybrid optimal descriptor based on the correlation weights derived from SMILES and molecular hydrogen-suppression graphs (HSG). They also performed validation by using the criterion index of ideality correlation (IIC). In this descriptor, a balance of the desirability correlation index (TF_2) was used to

develop the models. The experimental dataset was split indiscriminately into training, stealth training, calibration (~74%) and validation (~26%) sets. Four models were developed from the four splits, all of which were statistically satisfactory and stable.

Lotfi et al. [45] investigated the melting points of imidazolium-based ionic liquids using a QSPR approach to develop a melting point model for predicting the melting points of imidazolium-based ionic liquid datasets. A robust QSPR model was developed by applying the Monte Carlo algorithm of CORAL software to calculate the melting point values of 353 imidazole-like ionic liquids. Using a combination of SMILES and hydrogen-suppression molecular graphs (HSG), hybrid optimal descriptors were calculated and used to generate the QSPR model. Internal and external validation parameters were also used to assess the predictiveness and reliability of the QSPR models. Four slices were prepared from the dataset, each randomly assigned to four sets, namely the training set (~33%), the invisible training set (~31%), the calibration set (~16%) and the validation set (~20%). In the QSPR modeling, the values of various statistical features of the validation set, such as $R^2_{\text{Validation}}$, $Q^2_{\text{Validation}}$ and $\text{IIC}_{\text{Validation}}$, were found to be in the range of 0.7846–0.8535, 0.7687–0.8423 and 0.7424–0.8982, respectively. For mechanistic interpretation, they also extracted the structural properties that lead to an increase/decrease in melting point.

Makarov et al. [46] have also carried out some research on the melting point of ionic liquids. They developed a new model based on the SMILES translator and neural network, which showed a significant improvement in prediction accuracy compared to the previous studies. The model had $R^2 = 0.67$ and $\text{RMSE} = 44$ °C. The model is applicable to any type of ILs.

The ability to quantitatively predict ionic liquid (IL) properties using QSPR models is of great importance. It is therefore necessary to understand which modern machine learning (ML) methods combined with which types of molecular characterization are more suitable for this purpose. To address this issue, Baskin et al. [47] conducted a large-scale benchmarking study of QSPR models that were used to predict six important physical properties of ILs (density, conductivity, melting point, refractive index, surface properties) by combining three traditional ML methods and neural networks with seven different structures with five types of molecular representations (in the form of numerical molecular descriptors or SMILES text strings), melting point, refractive index, surface tension and viscosity. QSPR models for predicting the properties of ILs at eight different temperatures were developed using a multitask learning approach. The optimal combination of ML methods and molecular representation was determined for each property. A unified ranking system was introduced. The different ML methods and molecular representations were prioritized. This study shows that, on average, (i) nonlinear ML methods perform much better than linear methods, (ii) neural networks perform better than traditional ML methods and (iii) transformers, which are actively used in natural language processing (NLP), perform better than other types of neural networks due to the advanced ability to analyze chemical structures of ILs encoded into SMILES text strings. It has also employed a special “composition judgment” cross-validation scheme to assess how

much the predictive performance deteriorates for ILs consisting of cations and anions that are not present in the dataset.

12.3.4 Toxicity of Ionic Liquids and Their Mixtures

In recent years, ionic liquids (ILs) have attracted a great deal of attention due to their remarkable physicochemical properties. Despite the advantages of ILs, these compounds can cause persistent pollution and pose an environmental risk.

Ghaedi [48] used CORrelation And Logic (CORAL) software and cytotoxicity data for 225 ionic liquids to build QSAR models, where molecular structures are represented by SMILES symbols. These global SMILES descriptors account for the presence of a number of chemical elements and various types of chemical bonds (double bonds, triple bonds and stereochemistry). The balance of correlations (BC) of QSAR was constructed and compared with the classical scheme. The results of the three stochastic splits show that the R^2 for the reliable model predicting the external test set and Q^2 for the cross-validation range from 0.7315 to 0.8760 and 0.7062 to 0.8490, respectively. The optimal predictions obtained from the classical scheme are incorporated into the modeling process together with the global SMILES descriptors. The mean statistical characteristics of the external test set were as follows: $n = 44$, $R^2 = 0.8760$, $Q^2 = 0.8540$, standard error (s) = 0.529, mean absolute error (MAE) = 0.400 and Fischer F-ratio (F) = 297. The results indicate that the classical scheme is in terms of predictability of the QSAR model compared with the BC method. The results showed that the classical scheme was improved in terms of the predictability of the QSAR model compared with the BC method.

Lotfi et al. [49] predicted the minimum inhibitory concentration (MIC) of 204 of these ILs against *Staphylococcus aureus* (*S. aureus*) and the minimum bactericidal concentration (MBC) of 114 ILs using a QSAR based on a Monte Carlo approach. The molecular structures of all ILs are shown using the SMILES notation. For modeling pMIC and pMBC, a hybrid optimal descriptor was used, which was obtained by combining molecular maps and SMILES. For pMIC, the hybrid optimal descriptor was calculated by combining SMILES and a hydrogen-suppression molecular graph (HSG), while for pMBC the hybrid optimal descriptor was calculated by combining SMILES and a hydrogen filling graph (HFG). The full dataset was randomly divided into the training set, invisible training set, calibration set and validation set. QSAR models of pMIC and pMBC for ILs were developed by statistical analysis, and the index of correlation (IIC) was used as a benchmark for the predictive potential of these models. Their R^2 values for the training, invisible training, calibration and validation components were 0.8585–0.8853, 0.8523–0.8898, 0.8809–0.9240 and 0.8036–0.8903 for pMIC and 0.8357–0.8991, 0.8223–0.9306 for pMBC, respectively. The results indicate that the predictability of the QSAR model developed for all splits is at a high level. The method is shown to have reasonable predictive potential and mechanistic interpretation.

Ahmadi et al. [50] estimated the logarithm of the half-maximal effective concentration ($\log EC_{50}$) for the toxicity of ILs to the leukemic rat cell line IPC-81 based on a QSAR using a Monte Carlo approach with CORAL software. QSAR models were developed using mixed optimal descriptors for 304 different ionic liquids, including ammonium, imidazole, morpholine, phosphorus, piperidine, pyridine, pyrrolidine, quinoline, sulfate and plasmalogen ionic liquids. The SMILES notation of the ionic liquids was used to calculate the descriptor correlation weights (DCW). Four splits were performed from the entire dataset, and each split was randomly divided into four groups (training subset and validation set). The index of correlation (IIC) was used to assess the veracity and stability of the QSAR model. One of the QSAR models with statistical parameters of $R^2 = 0.85$, $CCC = 0.92$, $Q^2 = 0.84$ and $MAE = 0.25$ for the optimal split validation set was considered as a primary model.

12.4 Limitations and Outlook in Safety and Environmental Applications

12.4.1 Limitations

(1) Limited data

Biological systems are complex and have many indicators to measure toxicity. Besides, the toxicity data are few. Moreover, the physicochemical parameters of the nanomaterials are still unclear, which hindered the application of SMILES descriptors. Therefore, the nano-QSAR system needs to be tested, improved and refined.

(2) Limited descriptors

Descriptors largely determine the QSAR model. At present, there are very limited descriptors available for QSAR studies in nanomaterials. The predictive performance of models with different descriptors varies considerably.

(3) Unclear molecular mechanism

The toxic mechanism is very complex and not well understood, which needs to be explored from both experimental research and nano-QSAR research. From the existing studies, it can be found that mechanistic research is mainly focused on a few nano-metal oxides. Moreover, it is difficult to speculate on the molecular mechanism as the diverse research methods of experimental methods and standards.

(4) Insufficient database for model validation

Few studies meet the requirements of the OECD for QSAR models to calculate the application areas and explain the mechanisms. They commonly focused on the construction of predictive models without validating and giving comprehensive explanations.

12.4.2 Outlook

(1) Reliable experimental database

Currently, there are limited data available for nano-QSAR studies of metal oxide cytotoxicity. The reliability of the available experimental data on the biological effects of nanomaterials is yet to be verified due to the differences in their experimental methods and conditions. Therefore, the construction of a more complete and reliable nanomaterial cytotoxicity experimental database is still an important issue that needs to be addressed.

(2) More descriptors

Various nanomaterials have different compositions and different physicochemical properties. How to effectively characterize their structural and physicochemical characteristics is one of the key issues to be solved in nano-QSAR research. It is necessary to develop a series of new structural descriptors, graphical descriptors and other molecular descriptors to effectively characterize and describe their nanostructures, so as to establish more accurate and reliable prediction models.

(3) Mechanistic explanation

The mechanisms of toxicity of nano-metal oxides are complex. Although much research has been carried out, the underlying mechanism of toxicity of nano-metal oxides needs further in-depth research. This will provide guidance for the safe design, synthesis and application of nanoparticles.

(4) Flammability and toxicity of ionic liquids

SMILES has shown excellent performance in the field of ionic liquids. There are still many directions to be developed for the research on the flammability and toxicity of ionic liquids.

References

1. Katritzky AR, Lobanov VS, Karelson M (1995) *Chem Soc Rev* 24(4):279–287. <https://doi.org/10.1039/cs9952400279>
2. Katritzky AR, Maran U, Lobanov VS, Karelson M (2000) *J Chem Inf Comput Sci* 40(1):1–18. <https://doi.org/10.1021/ci9903206>
3. Katritzky AR, Perumal S, Petrukhin R, Kleinpeter E (2001) *J Chem Inf Comput Sci* 41(3):569–574. <https://doi.org/10.1021/ci000099t>
4. Katritzky AR, Fara DC (2005) *Energ Fuel* 19(3):922–935. <https://doi.org/10.1515/znb-2006-0403>
5. Taskinen J, Yliruusi J (2003) *Adv Drug Deliver Rev* 55(9):1163–1183. [https://doi.org/10.1016/s0169-409x\(03\)00117-0](https://doi.org/10.1016/s0169-409x(03)00117-0)
6. Yaffe DL (2001) A neural network approach for estimating physicochemical properties using quantitative structure-property relationships (QSPRs). Dissertation, University of California, Los Angeles
7. Mattioni BE (2003) The development of quantitative structure-activity relationship models for physical property and biological activity prediction of organic compounds. Dissertation, The Pennsylvania State University

8. Estrada E, Molina E (2001) *J Chem Inf Comput Sci* 41(3):791–797. <https://doi.org/10.1021/ci000156i>
9. Karelson M, Lobanov VS, Katritzky AR (1996) *Chem Rev* 96(3):1027–1044. <https://doi.org/10.1021/cr950202r>
10. Draper NR, Smith H (1998) *Applied regression analysis*. Wiley, New York
11. Miller A (2002) *Subset selection in regression*. Chapman and Hall/CRC, New York
12. Mitchell BE, Jurs PC (1997) *J Chem Inf Comp Sci* 37(3):538–547. <https://doi.org/10.1021/ci960175i>
13. Holland J (1975) *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor
14. Rogers D, Hopfinger AJ (1994) *J Chem Inf Comput Sci* 34(4):854–866. <https://doi.org/10.1021/ci00020a020>
15. Svante W, Sjöström M, Eriksson L (2001) *Chemometr Intell Lab* 58(2):109–130. [https://doi.org/10.1016/S0169-7439\(01\)00155-1](https://doi.org/10.1016/S0169-7439(01)00155-1)
16. Vapnik V (1999) *The nature of statistical learning theory*. Springer Science & Business Media
17. Vapnik VN (1999) *IEEE Trans Neural Netw* 10(5):988–999. <https://doi.org/10.1109/72.788640>
18. Gunn SR, Brown M, Bossley KM (1997) In: *International symposium on intelligent data analysis*, Aug 1997. Springer, Heidelberg, pp 313–323
19. Weininger D, Weininger A, Weininger JL (2002) *J Chem Inf Comput Sci* 29(2):97–101. <https://doi.org/10.1021/ci00062a008>
20. Buzea C, Pacheco I, Robbie K (2007) *Biointerphases* 2(4):MR17–MR71. <http://doi.org/10.1116/1.2815690>
21. Service RF (2003) *Science* 300(5617):243. <https://doi.org/10.1126/science.300.5617.243a>
22. Brumfiel G (2003) *Nature* 424(6946):246–249. <https://doi.org/10.1038/424246a>
23. Dowling AP (2004) *Mater Today* 7(12):30–35. [https://doi.org/10.1016/s1369-7021\(04\)00628-5](https://doi.org/10.1016/s1369-7021(04)00628-5)
24. Masciangioli T, Zhang W-X (2003) *Environ Sci Technol* 37:102A–108A. <https://doi.org/10.1021/es0323998>
25. Toropov AA, Benfenati E (2007) *Eur J Med Chem* 42(5):606–613. <https://doi.org/10.1016/j.ejmech.2006.11.018>
26. Toropov AA, Rasulev BF, Leszczynska D, Leszczynski J (2008) *Chem Phys Lett* 457(4–6):332–336. <https://doi.org/10.1016/j.cplett.2008.04.013>
27. Toropov AA, Toropova AP, Benfenati E, Gini G, Puzyn T, Leszczynska D, Leszczynski J (2012) *Chemosphere* 89(9):1098–1102. <https://doi.org/10.1016/j.chemosphere.2012.05.077>
28. Toropova AP, Toropov AA (2013) *Chemosphere* 93(10):2650–2655. <https://doi.org/10.1016/j.chemosphere.2013.09.089>
29. Toropova AP, Toropov AA, Manganelli S, Leone C, Baderna D, Benfenati E, Fanelli R (2016) *NanoImpact* 1:60–64. <https://doi.org/10.1016/j.impact.2016.04.003>
30. Manganelli S, Leone C, Toropov AA, Toropova AP, Benfenati E (2016) *Chemosphere* 144:995–1001. <https://doi.org/10.1016/j.chemosphere.2015.09.086>
31. Toropov AA, Toropova AP (2021) *Sci Total Environ* 772:145532. <https://doi.org/10.1016/j.scitotenv.2021.145532>
32. Toropova AP, Toropov AA, Rallo R, Leszczynska D, Leszczynski J (2015) *Ecotoxicol Environ Saf* 112:39–45. <https://doi.org/10.1016/j.ecoenv.2014.10.003>
33. Toropova AP, Toropov AA, Benfenati E, Korenstein R, Leszczynska D, Leszczynski J (2015) *Environ Sci Pollut Res Int* 22(1):745–757. <https://doi.org/10.1007/s11356-014-3566-4>
34. Pan Y, Li T, Cheng J, Telesca D, Zink JI, Jiang J (2016) *RSC Adv* 6(31):25766–25775. <https://doi.org/10.1039/c6ra01298a>
35. Choi JS, Trinh TX, Yoon TH, Kim J, Byun HG (2019) *Chemosphere* 217:243–249. <https://doi.org/10.1016/j.chemosphere.2018.11.014>
36. Cao J, Pan Y, Jiang Y, Qi R, Yuan B, Jia Z, Jiang J, Wang Q (2020) *Green Chem* 22(11):3512–3521. <https://doi.org/10.1039/d0gc00933d>
37. Ahmadi S (2020) *Chemosphere* 242:125–192. <http://doi.org/10.1016/j.chemosphere.2019.125192>

38. Toropova AP, Toropov AA, Leszczynski J, Sizochenko N (2021) *Environ Toxicol Pharmacol* 86:103665. <http://doi.org/10.1016/j.etap.2021.103665>
39. Toropova AP, Toropov AA (2022) *Environ Technol* 1–8. <http://doi.org/10.1080/09593330.2022.2093655>
40. Saldana DA, Starck L, Mougin P, Rousseau B, Pidol L, Jeuland N, Creton B (2011) *Energ Fuel* 25(9):3900–3908. <https://doi.org/10.1021/ef200795j>
41. Toropova AP, Toropov AA, Leszczynska D, Leszczynski J (2020) *New J Chem* 44(12):4858–4868. <https://doi.org/10.1039/d0nj00121j>
42. Toropova AP, Toropov AA, Carnesecchi E, Benfenati E, Dorne JL (2019) *Chem Pap* 74(2):601–609. <https://doi.org/10.1007/s11696-019-00903-w>
43. Gantzer P, Creton B, Nieto-Draghi C (2021) *J Chem Inf Model* 61(9):4245–4258. <https://doi.org/10.1021/acs.jcim.1c00803>
44. Lotfi S, Ahmadi S, Kumar P (2021) *J Mol Liq* 338:116465–116472. <https://doi.org/10.1016/j.molliq.2021.116465>
45. Lotfi S, Ahmadi S, Kumar P (2021) *RSC Adv* 11(54):33849–33857. <https://doi.org/10.1039/d1ra06861j>
46. Makarov DM, Fadeeva YA, Shmukler LE, Tetko IV (2021) *J Mol Liq* 344:117722. <https://doi.org/10.1016/j.molliq.2021.117722>
47. Baskin I, Epshtein A, Ein-Eli Y (2022) *J Mol Liq* 351:118616. <https://doi.org/10.1016/j.molliq.2022.118616>
48. Ghaedi A (2015) *J Mol Liq* 208:269–279. <https://doi.org/10.1016/j.molliq.2015.04.049>
49. Lotfi S, Ahmadi S, Zohrabi P (2020) *Struct Chem* 31(6):2257–2270. <https://doi.org/10.1007/s11224-020-01568-y>
50. Ahmadi S, Lotfi S, Kumar P (2022) *Toxicol Mech Method* 32(4):302–312. <https://doi.org/10.1080/15376516.2021.2000686>

Chapter 13

SMILES and Quasi-SMILES in QSAR Modeling for Prediction of Physicochemical and Biochemical Properties



Siyun Yang, Supratik Kar, and Jerzy Leszczynski

Abstract QSAR modeling of diverse physicochemical and biochemical properties of organic chemicals and nanomaterials utilizing the simplified molecular-input line-entry system (SMILES) and quasi-SMILES representation is quite a popular approach nowadays. Along with the SMILES, the quasi-SMILES approach offers the likelihood to identify and weigh the statistical importance of various eclectic data accessible for computational systematization and analysis. Therefore, the quasi-SMILES can be helpful as a tool for drug design, environmental risk assessment, and regulation caused by applying nanomaterials and organic chemicals as the method gives the possibility to consider building up corresponding models. The Monte Carlo method is applied to build up the QSAR modeling employing information collected from SMILES and quasi-SMILES. The model can be freely developed using open-access CORrelation And Logic (CORAL) software. The quasi-SMILES is an ideal approach for complex chemical systems like nanomaterials where there is no limitation to choose the list of eclectic data to make a reliable, efficient, and predictive QSAR model. In the present book chapter, we will talk about the fundamental of SMILES and quasi-SMILES-based QSAR models and their major applications in physicochemical and biochemical properties prediction.

Keywords CCI · IIC · SMILES · Monte Carlo · QSAR · Quasi-SMILES

Abbreviations

ΔG^\ddagger Gibb's activation free energy

S. Yang · S. Kar (✉)

Chemometrics and Molecular Modeling Laboratory, Department of Chemistry, Kean University, 1000 Morris Avenue, Union, NJ 07083, USA
e-mail: skar@kean.edu

J. Leszczynski

Department of Chemistry, Physics and Atmospheric Sciences, Interdisciplinary Center for Nanotoxicity, Jackson State University, Jackson, MS 39217, USA

CCI	Correlation Contradiction Index
GNPs	Gold nanoparticles
HSG	Hydrogen-suppressed graphs
IIC	Index of ideality of correlation
MOFs	Metal–organic frameworks
QSAR	Quantitative structure–activity relationship
QSGFEAR	Gibb’s free energy of activation relationship
QSPRs	Quantitative structure–property relationships
QSRR	Quantitative structure–retention relationship
QSTR	Quantitative structure–toxicity relationship
SADT	Self-accelerating decomposition temperature
SMILES	Simplified molecular-input line-entry system
WS	Water solubility

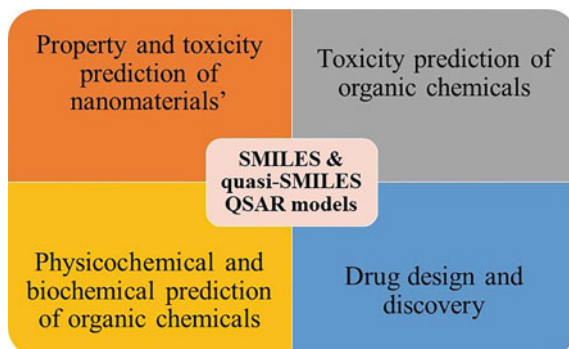
13.1 Introduction

Simplified molecular-input line-entry system (SMILES) and quasi-SMILES is a series of representative symbols including all accessible information from the dataset, like the structure of molecules, physicochemical conditions of the molecule, size of nanomaterials, etc. [1]. Among the major *in silico* approaches, quantitative structure–activity/toxicity/property relationships (QSARs/QSTRs/QSPRs) can utilize limited experimental resources and need minimal computing time, saving money. QSAR modeling can deliver significant information at a low expense for drug discovery and development by facilitating rational strategy design. In addition, the QSAR approach can predict the chemical response of a relatively large number of compounds within the chemical domain using the response data of a small number of chemicals which is commonly used in predictive toxicology studies for the evaluation of chemical risks [2].

Due to the convenience that SMILES and quasi-SMILES brought, modeling with these notations has become increasingly popular among scientists. The first and foremost reason is easy to represent any molecules followed by features calculation for modeling any physicochemical and biochemical properties. Quasi-SMILES is an analogy of traditional SMILES which contain some additional information besides the molecular architecture [3]. To develop the QSAR models, Toropova et al. [4] had developed CORAL software (<http://www.insilico.eu/coral>) where 2D-optimal descriptors can be calculated with so-called correlation weights for attributes of SMILES and quasi-SMILES where the correlation weights are obtained as results of the unique Monte Carlo optimization [5]. Although, additional features and conditions may need to be considered during modeling to develop predictive QSAR models.

A series of physicochemical and biochemical properties were already modeled using SMILES and quasi-SMILES employing Monte Carlo approach using CORAL

Fig. 13.1 Major research areas of SMILES and quasi-SMILES-based QSAR model



software. Toropov and Toropova [6] also proposed an index of ideality of correlation (IIC), which has been tested to improve the predictive potential of diverse QSAR endpoints. The fundamental aim of the IIC is to unite sensitivity to correlation, dispersion, and symmetry of the distribution of images around the diagonal. Another important index, Correlation Contradiction Index (CCI), has been proposed by Toropov and Toropova [7] as a criterion of predictive potential. Therefore, the whole modeling process is simple, as no 3D structure is required for the study. The entire model can be developed in CORAL software followed by strong predictive indices like IIC and CCI.

SMILES and quasi-SMILES-based method was successfully employed for the development of model for mutagenicity and mutagenic potential of fullerenes and multi-walled carbon nanotubes [8, 9], toxicity of nanoparticles [10, 11], and cytotoxicity of metal oxide nanoparticles to bacteria *Escherichia coli* [12], predict behavior of complex systems like peptides [13, 14], physicochemical [15, 16] and biochemical properties of polymers [17]. The wide range of successful applications of these mentioned approaches makes it one of the most powerful prediction tools (Fig. 13.1).

13.2 Fundamentals of SMILES and Quasi-SMILES

Simplified molecular-input line-entry system (SMILES) is a chemical notation that lets a user depict a chemical structure in a way the computer system can utilize. SMILES is a quickly learned and flexible notation that allows for a simple representation of any molecular structure. There are defined equivalences between the representation of the molecular structure by graphs and using SMILES approach [18].

To model diverse endpoints, 0D to 7D descriptors have evolved over the years [19]. But it's always best to use simple descriptors from 0D to 2D, which are easy to compute and interpret the developed QSAR models [2]. Optimal descriptors have been developed and refined along with advances in QSAR approaches. Initially, the molecular graph-derived features or descriptors were the basis for building a QSAR

model. A similar idea has been introduced and developed for SMILES and SMILES attributes. It can be summed up as follows [20]:

- (a) Each SMILES of the modeling set computes a list of attributes, x_{kj} :

$$\text{SMILES}_k \rightarrow \{x_{k1}, x_{k2}, \dots, x_{km}\}. \quad (13.1)$$

- (b) Followed by the Monte Carlo method offers correlation weights for the total set of attributes. They are extracted from all SMILES notations of the modeling set, which provide the maximal correlation coefficient between the studied endpoint and sums of correlation weights for SMILES of the modeling set:

$$\text{Monte Carlo method} \rightarrow \{\text{CW}(x_{k1}), \text{CW}(x_{k2}), \dots, \text{CW}(x_{km})\}. \quad (13.2)$$

- (c) A one-variable linear equation represents the predictive model:

$$\text{EP}_k = C_0 + C_1 \times \sum_{x_{kj} \in \text{SMILES}} \text{CW}(x_{kj}) = C_0 + C_1 \times \text{DCW}(T^*, N^*). \quad (13.3)$$

In the vector and matrix depictions, this approach can be explained as the following:

$$\begin{pmatrix} \text{MS}_1 \\ \text{MS}_2 \\ \dots \\ \text{MS}_n \end{pmatrix} \rightarrow \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \leftrightarrow \begin{pmatrix} E_1 \\ E_2 \\ \dots \\ E_n \end{pmatrix}, \quad (13.4)$$

where MS_k are molecular structures available from SMILES or graphs and x_{kj} illustrate molecular features extracted from SMILES, while the basis of preparing quasi-SMILES can be removed from a graph, SMILES, and eclectic data.

The traditional approach assumes that an endpoint depends on the molecular structure. However, there are cases in which this approach has to be revised. There are also situations where one can expect that the endpoint depends on other conditions (concentration, temperature, dose, etc.) and circumstances (magnetic field, the presence/absence of illumination, different times of exposure, etc.). In this case, instead of the hypothesis: "Endpoint (Y) = function (Molecular Structure)," one can consider the following hypothesis: "Endpoint = functions (Eclectic Data)."

$$\begin{pmatrix} \text{ED}_1 \\ \text{ED}_2 \\ \dots \\ \text{ED}_n \end{pmatrix} \rightarrow \begin{bmatrix} \text{CW}(x_{11}) & \text{CW}(x_{12}) & \dots & \text{CW}(x_{1m}) \\ \text{CW}(x_{21}) & \text{CW}(x_{22}) & \dots & \text{CW}(x_{2m}) \\ \dots & \dots & \dots & \dots \\ \text{CW}(x_{n1}) & \text{CW}(x_{n2}) & \dots & \text{CW}(x_{nm}) \end{bmatrix} \leftrightarrow \begin{pmatrix} E_1 \\ E_2 \\ \dots \\ E_n \end{pmatrix} \quad (13.5)$$

ED_k can be defined as symbols correlation weights obtained from quasi-SMILES, $\text{CW}(x_{kj})$ is obtained experimental data for the endpoint, E_k . Finally, the vector

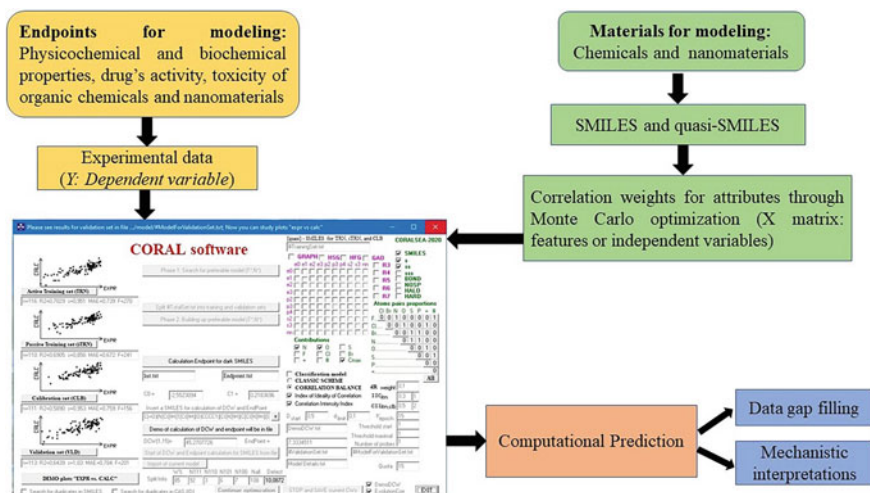


Fig. 13.2 General scheme of SMILES and quasi-SMILES QSAR modeling

computed from eclectic data signifies quasi-SMILES. Interesting to point out that, like SMILES, quasi-SMILES isn't inevitably the depiction of molecular features.

Once the user computed Eq. 13.5, the Monte Carlo method will be utilized to optimize the correlation weights. The explained methodology defines the mechanical interpretation of the model based on the correlation weights of effective features obtained from quasi-SMILES. Having the numerical data on the correlation weights of features that takes place in several runs of the Monte Carlo optimization, one can extract three categories of these features:

1. Features with negative values of the correlation weight in all runs, which are reasons for endpoint decrease.
 2. Features with positive values of the correlation weight in all runs, which are reasons for endpoint increase.
 3. Features with both positive and negative values of the correlation weight in different runs of the optimization, which are features with an unclear role.
- A complete flow diagram of SMILES and quasi-SMILES QSAR model is illustrated in Fig. 13.2.

13.3 Application of SMILES and Quasi-SMILES-Based QSAR Model

To better understand readers, we have divided multiple physicochemical and biochemical properties into diverse materials, properties and toxicity.

13.3.1 Nanoparticles Toxicity and Property Prediction

Quasi-SMILES could act as a flexible foundation for accessing the regulation and environmental risk of nano-QSAR [21]. The technique served as a bridge between experimentalists and model developers for nanomaterials-related endpoints. The boundary rejection between the effect of the biochemical reality of molecular level substance and the experiment conditions effect at the macro-level permits the development of models that are epistemologically more reliable than traditional ways. The reason is solely based on the interdependence between molecular structure and biological activity (without taking into account experimental conditions). Nanoparticle physicochemical and biochemical behavior models are required for developing and applying new industrial accomplishments like food, makeup, and medicine without detrimental impacts on the environment and human health.

Nano-QSPR/QSAR should always follow the five OECD principles. In addition, it may be necessary to specify new regulations for nano-QSPR/QSAR that represent the nano-nature of the compounds under study. For example, the principles should consider the experimental settings and the quality of the applicable equipment. In this case, the software could access environmental regulation and risk assessment. Nanomaterials exhibit unique physicochemical and biological properties. The logic of nanoparticles differs from the logic governing the behavior of conventional substances. An apparent distinction between nano-phenomena and phenomena associated with traditional substances was the vast number of physicochemical circumstances that interact and mutually influence one another and the difficulty in identifying the nature of these relationships. The quasi-SMILES method allows for detecting and evaluating the statistical significance of various eclectic data accessible for computer systematization and analysis. Moreover, the approach allows for the relatively rapid modification of computational experiment bases (adding or removing eclectic conditions or circumstances).

The quasi-SMILES technique could be utilized as a regulatory and environmental risk assessment tool resulting from nanomaterials since the approach allows for the incorporation of the essential properties of the molecular structure and the experimental settings.

Toropov and Toropova [7] reported that they have successfully applied the quasi-SMILES to predict the mutagenicity of silver nanoparticles under different conditions. With the 72 data points, the data was equally distributed into training, invisible training, calibration, and validation group, and the calculation is performed 15 times. As a result, two target functions were optimized, TF_1 and TF_2 . Based on the rule of random effect of QSAR, fifteen random splits were completed with both functions and indicated that TF_2 had better performance, as shown below:

$$N_{cp} = -7.240(\pm 5.835) + 26.43(\pm 2.92) \times DCW(1, 10). \quad (13.6)$$

Additionally, the (i) Index of Ideality of Correlation (IIC) and (ii) Correlation Contradiction Index (CCI) were calculated based on TF_2 ; the result showed that IIC

has a value of R equal to 0.73, and the CCI showed a value of 0.78 which was better in comparison. The experimental result demonstrated that Quasi-SMILES could be a predictive model for silver nanoparticle mutagenicity. Simultaneously, IIC and CCI could be critical models to examine models' predictive potential.

Another application was executed on gold nanoparticles in 2021 [22]. A549 cell uptake potential of gold nanoparticles (GNPs) model under different conditions was computed, and Monte Carlo method was used for optimization. In this case, quasi-SMILES was defined as an information system with fragments about the phenomena of the inhibitory activity of GNPs under defined conditions. From the original target function, four more target functions were optimized with the criteria of IIC and CII below:

$$TF_0 = r_A + r_B - |r_A - r_P| \times 0.1 \quad (13.7)$$

$$TF_1 = TF_0 + IIC_C \times 0.5 \quad (13.8)$$

$$TF_2 = TF_0 + IIC_C \times 0.5 + IIC_P \times 0.5 \quad (13.9)$$

$$TF_3 = TF_0 + CII_C \times 0.5 \quad (13.10)$$

$$TF_4 = TF_0 + CII_C \times 0.5 + CII_P \times 0.5. \quad (13.11)$$

All four target functions that were used to compute models for cellular absorption of GNPs can predict the cell uptake. The created models enable mechanistic interpretation and promoters of an increase or decrease of the investigated endpoint to be identified. The use of the CII values for both the passive training set and the calibration set was what gives the model with the best predictive potential that has been seen in the case of the target function.

Quasi-SMILES could work as a foundation of nanoparticle toxicity and risk assessment. In this experiment, quasi-SMILES was a series of symbols that serve as codes for the settings of studies designed to evaluate the toxicity of ZnO nanoparticles to rats when injected intraperitoneally [23]. Correlation weights of each fragment from quasi-SMILES could be accessed by the Monte Carlo method and used to develop the variable models as per Eq. 13.12:

$$\text{Renal Factor} = C_0 + C_1 \times DCW(T, N). \quad (13.12)$$

The authors performed five random split sub-systems of training and validation. As a result, a 0.957 determination coefficient and a 7.25 root mean square error were gained. In this study, the applicability domain depended on the space of accessible qualities of quasi-SMILES, which corresponded to experimental conditions. If the experimental circumstances were not included in the list of experimental conditions, it becomes challenging to make a credible prediction of the endpoint using

the model applied here. The disclosed method could be used to produce predictions for the outcomes of intraperitoneal nanoparticle injections in rats, as well as for other experiments that can be represented by quasi-SMILES that were comparable to those.

In 2016, a model of an effective method for predicting the genotoxicity of carbon nanotubes was provided [24]. The experimental results of the bacterial reverse mutation test (TA100) on multi-walled carbon nanotubes (MWCNTs) were gathered from the published literature and analyzed as the last step. A mathematical model of the endpoint was developed using the optimum descriptors computed with the Monte Carlo approach. The model is a function of (i) dosage (g/plate), (ii) metabolic activation, and (iii) two kinds of MWCNTs. The method employed yielded a semi-quantitative prediction for three distinct distributions of experimental data: visible training and calibration sets and an invisible validation set. The predictive capability of these models varies. In the created models, quasi-SMILES exist with “atypical” behavior which suggests they are outliers even when included in the training set. However, deleting these quasi-SMILES conditions reduces the predictive capability of the models.

With the quasi-SMILES, the toxicity of *Daphnia magna* to nano-mixtures was also predictable [25]. As a mathematical function of experimental circumstances, toxicity is simulated. Nano-QSAR for predicting the toxicity of nano-mixtures was constructed utilizing a database of experimental data and the Monte Carlo method for optimization to calculate optimal descriptors with the potential predictive criteria CCI and IIC. The optimized target functions TF_1 and TF_2 were listed below:

$$TF_0 = r_{AT} + r_{PT} - |r_{AT} - r_{PT}| \times 0.1 \quad (13.13)$$

$$TF_1 = TF_0 + IIC_C \times 0.5 \quad (13.14)$$

$$TF_2 = TF_1 + CII_C \times 0.5. \quad (13.15)$$

The described quasi-SMILES method yields models of nano-mixtures toxicity of TiO_2 nanoparticles with high prediction ability. Compared to the IIC, the CCI is a more effective predictability criterion for nano-QSAR analysis as per the obtained outcome in the present study. The quasi-SMILES method can serve as the foundation for a language that facilitates communication between experimentalists and modelers of the properties or activity of nanomaterials.

Nano-QSPR model could also be modeled by quasi-SMILES which was proposed by Jafari et al. in 2022 [1]. Utilizing nanofluids as a suspension of nanoparticles in a common liquid was a relatively new subject that has attracted considerable interest recently. Quasi-SMILES is a series of representative symbols including all accessible information, like molecular structure and physicochemical conditions. This notation was used to illustrate the structure of nanofluids in consideration of the power of quasi-SMILES molecular representation to characterize diverse facts, such as nanoparticle size and form. To construct models, three random splits of

each dataset into active training, calibration, passive training, and validation sets were evaluated, and statistical assessment revealed that models generated using CII were superior to those developed using IIC. The following two target functions were examined via Monte Carlo optimization:

$$TF_0 = r_{AT} + r_{PT} - |r_{AT} - r_{PT}| \times 0.1 \quad (13.16)$$

$$TF_1 = TF_0 + IIC_C \times 0.5 \quad (13.17)$$

$$TF_2 = TF_1 + CII_C \times 0.5. \quad (13.18)$$

In these formulas, r_{AT} and r_{PT} were the experimental and anticipated values of the endpoint for the active training set and passive training set, respectively. Due to the unique uses of nanofluids, it was necessary to optimize nanofluids' composition and empirical circumstances rather than their intended thermophysical characteristics. The size of nanoparticles affects viscosity; thus, it was possible to estimate the model's outcome. Through the analysis, TF_2 was the best in the running datasets. It was determined that model creation based on the CII was statistically more trustworthy than model generation based on the IIC.

Metal oxide nanoparticles could be modeled by quasi-SMILES [26] for the risk assessment and safety evaluation which was typically a time-consuming and expensive experimentally. Hence, computational analyses were frequently employed to supplement actual testing. Structure–activity relationships (SAR) modeling was one of the most time-efficient approaches. The Sustainable Nanotechnology (S2NANO) collection comprises 574 experimental cell viability and toxicity for Al_2O_3 , CuO, Fe_2O_3 , Fe_2O_4 , SiO_2 , TiO_2 , and ZnO were included in the model construction settings. A quasi-SMILES molecular representation-based QSAR models were built up for classification and regression-based structure–activity relationship. The quasi-SMILES algorithm had all available data, including nanoparticle structural characteristics like molecular structure, core size, and relevant experimental factors like cell line, dose, exposure time, assay, hydrodynamic size, and surface charge. Regression models generated sufficient predictive ability. However, classification models displayed more precision. Incorporating both theoretical and experimental data into quasi-SMILES descriptors might be helpful for early risk evaluation of metal oxide nanoparticles. The proposed descriptors are easily calculable and might be utilized to create statistically sound models. Since the examined dataset contained measurements of cell viability and cytotoxicity under a range of experimental settings, seven types of nanoparticles were capable of capturing by the developed models and generalized safety pictures.

A novel method for constructing and evaluating predictive models of the octanol/water partition coefficient for gold nanoparticles was set up by Toropova and Toropov [27]. The partition coefficient for nanoparticles in octanol/water is an essential parameter for estimating the ecological destiny of these novel chemicals rapidly disseminating in everyday life. The validation of a model's prediction ability

is a crucial component of QSPRs. The so-called system of self-consistent models may provide a novel strategy for validating predictive capability. The measure of self-consistency is the mean of the correlation coefficients found for several models on distinct validation sets. The purpose of the study was to assess the adequacy of the self-consistency of models derived from two methods to identify a superior modeling method for octanol/water partition coefficients for gold nanoparticles (GNPs). The models mentioned above are predicated on the representation of GNPs by so-called quasi-SMILES, which are unique sequences of symbols that translate data about the architecture and operating circumstances of GNPs. Two optimized target functions are listed below.

$$TF_1 = r_{AT} + r_{PT} - |r_{AT} - r_{PT}| \times 0.1 \quad (13.19)$$

$$TF_2 = TF_1 + IIC_C \times 0.5. \quad (13.20)$$

The first method involves the Monte Carlo optimization of the correlation coefficient between the observed and anticipated outcomes. The second technique modifies the first by incorporating an extra criterion, IIC. The self-consistency and predictive capability of the second method are superior. Concurrently, it is demonstrated that the described quasi-SMILES approach yields a model of $\log P$ for gold nanoparticles that is highly resilient.

The models for solubility of fullerenes C[60] and C[70] were able to predict through SMILES and quasi-SMILES-based QSPR models [28]. Correlations of criteria of prediction ability of models for solubility of fullerenes C[60] and C[70] observed for the calibration (visible) set with determination coefficients of comparable models for validation sets (external, invisible). The IIC participated in the Monte Carlo optimization to establish a one-variable QSPR to forecast the solubility of fullerenes C[60] and C[70]. This significantly enhanced the forecasting capability of models for this solubility. Following a study of the statistical quality of the calibration set, better models may be selected based on the criteria of predictive potential, and the genesis of the potential predictive measures addressed is distinct. Two statistical elements, the correlation coefficient and the mean absolute error, are considered by the IIC, which may be a benefit.

13.3.2 Toxicity Predictions and Risk Assessment of Organic Chemicals

Quasi-SMILES could also be a helpful tool for removal rates of pharmaceuticals and dyes prediction in sewage [29]. In this study, quasi-SMILES codes could represent various eclectic conditions, such as the existence of light, X-rays beam impaction, and seasons. From the data collected from the literature, two CORAL models were constructed and stated below:

$$\text{Removal Rate (\%)} = -83.32 + 1.63 * \text{DCW}(1, 15) \quad (13.21)$$

$$\text{Removal Rate (\%)} = 16.61 + 0.589 * \text{DCW}(1, 20). \quad (13.22)$$

The approach described here gives quite efficient and predictive QSAR models. In addition, the process was much more straightforward with quasi-SMILES. Experimental methods provide more precise numerical data on removal rates, but predictive computer models are also required, at least for simple engineering decision estimations.

Based on Monte Carlo approach, organic compounds' ecotoxicological prediction toward *Pseudokirchneriella subcapitata* could be performed [30]. Acute toxicity was one of the most critical factors utilized in ecotoxicological risk assessment. *P. subcapitata* have been used in ecotoxicological investigations to determine the toxicity of several toxic compounds in freshwater. Using quantitative structure–toxicity relationship (QSTR) modeling, the toxicity of 334 distinct compounds on *P. subcapitata* was evaluated in terms of EC₁₀ and EC₅₀ values. Using CORAL software, the QSTR models were created by combining the target function (TF₂) and the IIC using a hybrid optimum descriptor generated from SMILES and molecular hydrogen-suppressed graphs (HSG). Overall, the approach of balancing of correlation with IIC was utilized to develop QSTR models. Using the IIC to create the QSTR models improved the robustness and predictability of the produced models, notably for the validation set. In addition, the generated QSTR models were nonparametric. Three random splits and four sets of single-split active training, invisible training, calibration, and validation sets were employed to prove the dependability of QSTR models.

The Monte Carlo method examines the adsorption affinity of azo dyes by applying new predicted statistical criteria [31]. Due to their chemical stability and simplicity of production, azo dyes are widely employed in several sectors. However, these colors are often recognized as hazardous environmental contaminants. Consequently, a mathematical model for the adsorption affinity of azo dyes may be used for medical and ecological problems. As a result of their chemical stability and simplicity of production, azo dyes are utilized in a variety of sectors. However, these colors are frequently recognized as significant environmental contaminants. Consequently, a mathematical model for the adsorption affinity of azo dyes may be used to solve problems in the fields of health and ecology. The optimal SMILES-based descriptors were used to create QSPR for the adsorption affinity of azo dyes to a substrate (DAF, kJ/mol) using the Monte Carlo approach. The IIC and the CII improved the model's predictive potential, primarily when they were used simultaneously.

$$\text{TF}_0 = r_{\text{AT}} + r_{\text{PT}} - |r_{\text{AT}} - r_{\text{PT}}| \times 0.1 \quad (13.23)$$

$$\text{TF}_1 = \text{TF}_0 + \text{IIC}_C \times 0.5 \quad (13.24)$$

$$TF_2 = TF_0 + CII_C \times 0.3 \quad (13.25)$$

$$TF_3 = TF_0 + CII_C \times 0.3 + IIC_C \times 0.5. \quad (13.26)$$

The IIC in TF_1 and CCI in TF_2 enhance the prediction capability of QSPR for DAF. The concurrent usage of these indices (TF_3) is particularly efficient. The significant absolute mean of the determination coefficient on ten random splits and the tiny dispersion of the value on ten random splits illustrate the benefit of the TF_3 .

The Monte Carlo approach for constructing models of the half-lives of hydrolysis of organic molecules was presented in 2021 [5]. The hydrolysis of organic molecules such as pesticides, pollutants, and pharmaceuticals can influence the destiny and behavior of environmental contaminants; thus, it is important to examine the substance's stability in water for various reasons. However, the actual measurements of all compounds would necessitate colossal resources, and computational models may become appealing. Using the CORAL program, QSPR models of hydrolysis were constructed. The 2D-optimal descriptor is computed using correlation weights for SMILES characteristics. The correlation weights are derived using a unique Monte Carlo optimization. The composition of five or six carbon rings is a crucial component of this strategy. The QSPR models for predicting the half-life of hydrolysis of organic compounds are based on the idea that "QSPR is a random occurrence." In other words, this strategy was evaluated using three random splits. In every instance, the CORAL program provides accurate models. Moreover, this method provides insights into the mechanism and has been validated using the external validation set. Once again, the unique and paradoxical capacity of the index of ideality of correlation (IIC) to increase the statistical quality of a model for the calibration and validation sets at the expense of the training set is verified.

SMILES could be used to develop a hybrid descriptor-based QSTR model for predicting the toxicity of dioxins and dioxin-like compounds using correlation intensity index and consensus modeling [32]. The study included 95 halogenated dioxins and relevant chemicals with endpoint pEC50 for developing 12 QSTR models based on the Monte Carlo algorithm in CORAL software. Three target functions were computed and optimized. CII was discovered to be a dependable indicator of the prediction ability of QSTR models. In terms of the promoter of increase or decrease for pEC50, the fragments responsible for the toxicity of dioxins and similar substances were also found. Four QSTR models were developed for each target function type to get accurate statistical findings. Conforming to the idea that "QSAR is a random event," three optimized functions were evaluated using four random splits.

Models for organophosphates compounds (OPC) binding to acetylcholinesterase (AChE) developed via representing the molecular structure were proposed by Toropova et al. [33]. QSARs are used to construct organophosphate prediction models. The determination coefficient for the validation set varied from 0.87 to 0.90, indicating that these models had a high predictive ability. These models were developed following the notion "QSAR is a random event," which states that the

predictive capacity of a method should be evaluated by dividing available data into training and test sets many times.

New robust and predictive models for AChE binding to OPC were developed. The sphere of applicability and a mechanistic explanation accompany these models. The statistical quality of the models investigated here is superior to that of models for the same endpoint generated by the CODESSA program [34]. The method [33] is reasonably straightforward and utilizes open-source CORAL software.

13.3.3 Miscellaneous Physicochemical and Biochemical Property Predictions of Organic Chemicals

13.3.3.1 Vapor Pressure (VP) Prediction

A self-consistent model system developed by Toropova et al. could be used to create and validate QSPRs [35]. The standard for these models' self-consistency is their ability to reproduce statistical quality despite variations in distributions. The model was built up by CORAL software:

$$\log \text{VP} = C_0 + C_1 \times \text{DCW}(T, N), \quad (13.27)$$

C_0 and C_1 stand for regression coefficient; DCW was the optimal descriptor calculated by SMILES. Monte Carlo method was performed for optimizations. Five splits of models were gained from the calculation:

$$\log \text{VP} = -3.838(\pm 0.012) + 0.2281(\pm 0.0010) \times \text{DCW}(3, 15) \quad (13.28)$$

$$\log \text{VP} = -3.941(\pm 0.012) + 0.2400(\pm 0.0010) \times \text{DCW}(3, 15) \quad (13.29)$$

$$\log \text{VP} = -3.625(\pm 0.011) + 0.2363(\pm 0.0009) \times \text{DCW}(3, 15) \quad (13.30)$$

$$\log \text{VP} = -3.708(\pm 0.011) + 0.2638(\pm 0.0009) \times \text{DCW}(3, 15) \quad (13.31)$$

$$\log \text{VP} = -4.241(\pm 0.010) + 0.1876(\pm 0.0010) \times \text{DCW}(3, 15). \quad (13.32)$$

All the target formulas have an R^2 of about 95%. One could assert that these models were generic and could be used for predictions since they were repeatable for the five splits, showing that they were not discovered by chance. Compared to various approaches, the system was applicable to realistically. The computational data confirmed that IIC could increase the predictive potential of the QSPR model. VP models were relatively simple to compute with SMILES structure.

13.3.3.2 Food Property Prediction

SMILES can be applied to food models. In 2019, Achary et al. proposed a model for sweetness [36]. With 315 molecules, QSAR models were built for the sweetness value ($\log S$). The descriptor used to build the model for $\log S$ was a hybrid optimal descriptor obtained by combining the following two descriptors: (1) molecular graph-based descriptor built from molecular feature correlation weights, (2) SMILES code describing the sweetener molecules. The 315-molecule dataset was partitioned into four random splits. The four QSAR models constructed for $\log S$ based on the IIC criterion were compared to four comparable models built using the “conventional approach” detailed elsewhere. The comparison found that IIC-built models had a superior statistical performance. The CORAL program could correctly simulate the sweetness potential ($\log S$). The IIC enables the statistical interpretation of CORAL-based QSAR models to be enhanced. The CORAL model had distinct criteria for estimating the quality of separating a given dataset into sets. Additionally, the requirements offered a statistically significant specification of the applicability domain (AD). The CORAL models’ statistical properties proved superior to the other models obtained from the 2D or 3D support vector regression.

13.3.3.3 Solubility Model

The water solubility (WS) model could be built up with SMILES introduced by Toropov et al. [37]. Water solubility models were constructed for 4224 molecules utilizing correlation weights of fragments of the SMILES, 2D graph invariants, and the ring hierarchy of the molecules. Two kinds of optimization were performed; one was the traditional version, and the other one was IIC version. Three splits were constructed for each version. The provided method produced reliable and resilient water solubility models. The IIC increased the descriptive models’ statistical quality. Despite the structural diversity of the examined compounds, the developed models were based on molecular structures without using 3D molecular descriptors, physicochemical descriptors, and/or quantum mechanical descriptors. The statistical quality of models derived using the IIC was equivalent to that of models constructed using recently proposed physicochemical endpoints and quantum mechanics descriptors.

The models for pesticide water solubility proposed in this publication are crucial from an ecological engineering standpoint [38]. Good in silico models were identified using the IIC of groups of QSPR models for the aqueous solubility of pesticides associated with the calibration sets. This comparison demonstrated that the high IIC set produces a model with superior statistical quality for the validation set. Even though there are extensive databases on solubility, the accurate prediction of the endpoint for novel compounds that might be used as pesticides is a crucial ecological challenge. The CORAL program provides a model for the WS of 1168 pesticides comparable to other solubility models proposed in the literature. Unfortunately, predictive models for various outcomes are susceptible to overtraining; the IIC aims to prevent or mitigate this. The IIC and correlation distribution enhance these models’ prediction

ability. This method compares a group of distinct data distributions into training and validation sets. Lastly, these models may be utilized for at least a preliminary mechanistic interpretation of specific molecular characteristics.

13.3.3.4 Self-accelerating Decomposition Temperature of Organic Peroxides Prediction

The breakdown of the organic peroxide is exothermic, and this heat can be employed for the polymers' or emulsion's intended or anticipated reactions. However, the unintended breakdown of these peroxides creates heat that is not efficiently dispersed and can lead to severe issues. Quasi-SMILES could be an appropriate way for computing self-accelerating decomposition temperature (SADT) [39]. A prediction model has been constructed with the help of IIC and the organic peroxides dataset. Every fragment or component of SMILES could be evaluated in terms of its incidence and statistical impact as a promoter of an increase or decrease in SADT. The benefit of dividing the SADT dataset into sets is an understandable criterion for generating robust CORAL-based QSPR models. However, the disadvantage of the SMILES-based technique, with or without the IIC criteria, is that it might take a long time to finish the optimization on huge datasets. When the SMILES attribute is not present in the molecular fragment, the CWs for such an attribute cannot be computed, resulting in a significant variation in optimal descriptors. CORAL-based QSPR models are novel models that appear to be sufficiently efficient for predicting critical features such as SADT and others.

13.3.3.5 Biological Activity of "Micelle-Polymer" Prediction

Modeling the biological activity of "micelle-polymer" samples with quasi-SMILES was created in 2018 [3]. The primary step of drug discovery is determining the molecular structure of novel pharmacological medicines. The delivery of active chemicals to the proper destinations within an organism must be clarified in detail. The polymeric structures identification serving as the foundation for transferring therapeutic substances into the body is one solution to the problem. Typically, models computed using the CORAL program offer mechanistic interpretation information regarding promoters of rise or reduction in several runs of an endpoint's optimization. There are only two fragments with multiple occurrences in training and calibration sets of quasi-SMILES with consistent positive correlation weights for arm star polymer and poly(ethylene glycol) methacrylate, and only one fragment with consistent negative correlation weights for Poly(ethylene glycol). However, the definition of quasi-SMILES and the strategy for extracting fragments of quasi-SMILES can be improved, for example, by separating micelles and polymers, defining not just digits but also integer coefficients, and possibly by making other changes. The stated investigation has demonstrated that suitable prediction models based on the provided

quasi-SMILES are theoretically conceivable. Quasi-SMILES is adequate representations of the micelle–polymer systems that allow for the construction of models. The technique utilized to define pieces of quasi-SMILES in the modeling process can be enhanced. Due to the high quality of the proposed models, further information on the physicochemical and biochemical properties of the micelle–polymer samples is not required.

13.3.3.6 CO₂ Uptake Prediction Model

Metal–organic frameworks MOFs were high-specific surface areas of hybrid organic–inorganic crystalline porous materials [40]. The model was examined and created that utilizes quasi-SMILES parameters such as Brunauer, Emmett, and Teller specific, surface area, pore volume, pressure, and temperature to MOFs for CO₂ uptake prediction. The dataset, which included 260 quasi-SMILES characteristics of MOFs, was randomly divided into training, validation, and test sets three times. Six QSPR models utilizing two target functions based on quasi-SMILES descriptors have been developed. The relevance of several eclectic characteristics of CO₂ increases and decrease ability of MOFs to absorb CO₂ is discussed:

$$\log(\text{CO}_2 \text{ uptake}) = C_0 + C_1 \times \text{DCW}(T^*, N^*) \quad (13.33)$$

$$\text{TF}_1 = R_{\text{TRN}} + R_{i\text{TRN}} - |R_{\text{TRN}} - R_{i\text{TRN}}| \times 0.1 \quad (13.34)$$

$$\text{TF}_2 = \text{TF}_1 + \text{IIC}_{\text{CAL}} \times W_{\text{IIC}}. \quad (13.35)$$

R_{TRN} and $R_{i\text{TRN}}$ experimental and projected $\log(\text{CO}_2 \text{ uptake})$ correlation coefficients for the training and invisible training sets, respectively. Optimization using Monte Carlo develops QSPR models based on IIC (TF_2). W_{IIC} was an empirical coefficient ($W_{\text{IIC}} = 0.2$ in this case), whereas IIC_{CAL} is the index of the ideality of correlation for the calibration set, which was defined by the calibration set's data.

The results show that TF_2 improves the predictability of models. Hence, simple and predictive models may be used to forecast the CO₂ capture capacity of MOFs. Based on the outputs of the QSPR models, the most critical factors that increase or decrease the CO₂ uptake capacity correspond with observations from experimental studies. According to the results of the QSPR model, the impacts of temperature and pressure on capturing CO₂ had been explored and are compatible with experimental observations. In addition, the model demonstrates that functionalization was a powerful technique for enhancing CO₂–MOF interaction and the CO₂ absorption of MOFs. According to the model interpretation results, the addition of basic N- and O-containing and double-bond-containing functional groups to the surfaces of organic linkers of MOFs was crucial for enhancing CO₂ absorption capabilities.

13.3.3.7 Monte Carlo Method-Based Gibbs Free Energy Studies

Construct quantitative structure under SMILES, Gibb's free energy of activation relationship (QSGFEAR) models with broad application and complete validation is feasible [41]. The experimental data of Gibb's activation free energy (ΔG^\ddagger) at seven different temperatures served as the endpoint, and the descriptor of correlation weight (DCW) was generated from the SMILES notation of the compounds. Two target functions were optimized in this case, one with CCI and one without.

$$TF_1 = R_{ATR\dot{N}} + R_{PTR\dot{N}} - |R_{ATR\dot{N}} - R_{PTR\dot{N}}| \times W_{IIC} \quad (13.36)$$

$$TF_2 = TF_1 + CII_{CAL} \times 0.3. \quad (13.37)$$

The QSGFEAR models were validated with a new statistical parameter called correlation intensity index (CII). A total of eight models were formed from the dataset of experimentally determined ΔG^\ddagger values, four using target function TF_1 ($W_{CII} = 0.0$) and four using target function TF_2 ($W_{CII} = 0.3$). It was found that the models built by applying CII were more accurate, robust, and consistent than those without CII. All the developed models were effective for predicting ΔG^\ddagger values reliably and consistently. The leading model was developed from split 3 using TF_2 with $R_{Val}^2 = 0.9108$. The mechanistic interpretation was done with the help of split 3, and the SMILES attributes responsible for the increase and decrease of ΔG^\ddagger value were identified.

Using the CII as a measure of predictors, a new target function was utilized to generate the SMILES-based descriptors. It was determined that the statistical quality of all the created models was adequate and that the developed model had an excellent predictive capacity. Examining the correlation weights of different molecular characteristics estimated through repeated Monte Carlo optimization runs provided a comprehensive mechanistic explanation of the increasing or decreasing structural features.

13.3.3.8 Glass Transition Temperature Studies

The optimal descriptors computed using SMILES indicated a structure of monomer units used to construct a model of the temperatures of glass transition of various polymers [42]. QSPRs were developed for the dataset mentioned above. Robust statistical quality characterizes the model of transition temperatures for glass. The molecular structure of matching monomers has been represented using SMILES. As the foundation for the one-variable model, the hybrid optimum descriptors generated using the so-called correlation weights of molecular characteristics taken from SMILES and molecular hydrogen-suppressed graph (HSG) were utilized. The IIC is a new criterion of the QSPR model's predictive ability. Here, the usefulness of the IIC as a tool to enhance the model's prediction capability for temperatures of glass transition

is demonstrated. The target function with a $R^2 = 0.90 \pm 0.01$ listed below:

$$\text{Tg}'K = C_0 + C_1 \times \text{DCW}(T^*, N^*). \quad (13.38)$$

The computation experiments conducted with three iterations of the Monte Carlo optimization verify that the predictive potential of models constructed with consideration of the IIC is acceptable, as the dispersion of the statistical quality of the models is satisfactory at 0.01 for the determination coefficient and 0.5 for the mean absolute error of the predicted glass transition temperatures.

13.3.3.9 Application on Chromatography Studies

QSRR of taste and fragrance compounds was investigated on a stationary phase methyl silicone OV-101 column utilizing correlation intensity index and consensus modeling by CORAL [43]. In chromatography, the QSRR is a critical technique for estimating unknown substances' retention period. Using the statistical parameter "correlation intensity index" (CII), the QSRR method is utilized to create robust models' of 1176 taste and aroma chemicals on the OV-101 glass capillary gas chromatographic column. QSRR models are constructed using the optimum descriptor, i.e., the descriptor of correlation weight (DCW) derived using SMILES notation. Using the balance of correlation technique, two target functions, TF_1 ($W_{\text{CII}} = 0$) and TF_2 ($W_{\text{CII}} = 0.3$), are used to create 12 QSRR models from six splits. According to statistical outcomes, models developed using CII perform better. The lists of structural characteristics responsible for variations in the retention index (RI) of tastes and scents compounds were retrieved as well. Utilizing the allocation structure of split 1 and the revised consensus, a consensus model is constructed (CM1). The test set's determination coefficient (R^2) for the modified consensus (CM1) model is calculated to be 0.9772, which is more than the leading model. QSRR models are statistically robust and validated with many validation parameters, exhibiting exceptional performance for external chemical prediction inside the AD.

Another relative research was proposed in 2022 [44]. A total of 1179 flavors and fragrances were included in this study for the creation of the QSRR model based on Monte Carlo algorithm in CORAL software. All organic molecules were encoded by SMILES notation to compute the correlation weight descriptor (DCW). The dataset of 1179 flavor and fragrance organic compounds was divided into nine subsets, each consisting of four subsets: training, invisible training, calibration, and validation. 18 QSRR models and two types of target functions were developed. The function of the index of ideality correlation (IIC) was thoroughly analyzed, and it was discovered that the QSRR models created by using the IIC were more robust and significant.

13.3.3.10 Models for Flammability of Binary Liquid Mixtures

The binary liquid mixtures QSPR model was also developed in 2020 [45]. Data on the flammability of binary liquid mixes is required for the categorization of liquid mixtures rationally. The list of related binary mixes with practical uses is extensive and is growing continuously. Therefore, accurate predictions of the endpoint might be advantageous. SMILES is the molecular structure representation. Quasi-SMILES is the extension of standard SMILES with the addition of symbols representing “eclectic” circumstances that might impact physicochemical endpoints. The application of quasi-SMILES to develop a model for the flammability of binary liquid mixtures revealed that the method provides an excellent model for the flash points ($^{\circ}\text{C}$) of binary organic mixtures.

The method enables the definition of a model’s mechanistic interpretation via a list of molecular characteristics that encourage flash points’ development (or reduction). The quasi-SMILES method yields relatively accurate predictions for the flash points of binary liquid mixes, including organic compounds. The IIC is an essential and valuable component of Monte Carlo optimization, as it provides the opportunity to enhance the prediction capability of models for flash points, for external, invisible validation sets. IIC is a new predictive capability metric. Successful attempts were made to utilize the IIC to enhance models for the flammability of binary liquid combinations.

13.3.3.11 Model for Disease Treatment Study

For a large database ($n = 141,706$), robust QSARs for hBACE-1 inhibitors (pIC_{50}) are developed [46]. New statistical criteria for evaluating the predictive capability of models are proposed and evaluated. These are the ideality of the correlation index (IIC) and the correlation intensity index (CII).

$$\text{TF}_1 = r_{\text{AT}} + r_{\text{PT}} - |r_{\text{AT}} - r_{\text{PT}}| \times 0.1 \quad (13.39)$$

$$\text{TF}_2 = \text{TF}_1 + \text{IIC}_C \times 0.5 \quad (13.40)$$

$$\text{TF}_3 = \text{TF}_2 + \text{CII}_C \times 0.5. \quad (13.41)$$

In conjunction with the Monte Carlo approach, the SMILES algorithm provides highly accurate models for hBACE-1 inhibitor action (IC_{50}). The best model after optimization is from TF_3 . Computational investigations with five distinct distributions for the active training set, passive training set, calibration set, and validation set demonstrated the statistical validity of these models. The models’ statistical properties for the validation set are assessed to measure the model’s predictive capability.

Despite the stochastic nature of the given technique, the proposed system of self-consistent models measures both the predictive potential of the applied approach (chosen model) and the repeatability of the findings.

13.4 Conclusion

The present chapter summarizes the major concepts of SMILES, quasi-SMILES, the Monte Carlo method, and Coral software and their application in diverse research field. SMILES and quasi-SMILES QSAR models have already been successfully applied on various endpoints. Due to the simplified notation, it is easier to build up a model for the aimed target. Using the Monte Carlo approach, CCI, and IIC parameters, one can make robust and significant QSAR models. From the existing models, SMILES and quasi-SMILES have satisfactory performance on environmental risk assessment, nanoparticle toxicity, property studies, drug design and discovery, environmental risk assessment, etc. The open-access CORAL software makes the whole modeling approach user-friendly and accessible to academics, industry, and independent researchers. One of this modeling method's unique features is modeling complex nanomaterial toxicity and properties using SMILES and quasi-SMILES followed by successful prediction. We believe this popular QSAR modeling approach will solve many unsolved queries of diverse scientific areas in the upcoming days.

Acknowledgements SY and SK want to thank the administration of Dorothy and George Hennings College of Science, Mathematics and Technology (HCSMT) of Kean University for providing research opportunities and resources.

Declaration of Competing Interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this chapter.

References

1. Jafari K, Fatemi MH, Toropova AP, Toropov AA (2022) Chemom Intell Lab Syst 222:104500. <https://doi.org/10.1016/j.chemolab.2022.104500>
2. Roy K, Kar S, Das RN (2015) A primer on QSAR/QSPR modeling: fundamental concepts. Springer
3. Toropov AA, Toropova AP, Benfenati E, Diomede L, Salmona M (2018) Struct Chem 29(4):1213–1223. <https://doi.org/10.1007/s11224-018-1115-3>
4. Toropova AP, Toropov AA, Benfenati E, Rallo R, Leszczyńska D, Leszczynski J (2017) Development of Monte Carlo approaches in support of environmental research
5. Toropov AA, Toropova AP, Lombardo A, Roncaglioni A, Lavado GJ, Benfenati E (2021) SAR QSAR Environ Res 32(6):463–471. <https://doi.org/10.1080/1062936x.2021.1914156>
6. Toropov AA, Toropova AP (2017) Mutat Res Genet Toxicol Environ Mutagen 819:31–37. <https://doi.org/10.1016/j.mrgentox.2017.05.008>

7. Toropov AA, Toropova AP (2019) *Sci Total Environ* 681:102–109. <https://doi.org/10.1016/j.scitotenv.2019.05.114>
8. Toropov AA, Toropova AP (2014) *Chemosphere* 104:262–264. <https://doi.org/10.1016/j.chemosphere.2013.10.079>
9. Toropov AA, Toropova AP (2015) *Chemosphere* 124:40–46. <https://doi.org/10.1016/j.chemosphere.2014.10.067>
10. Toropova AP, Toropov AA (2013) *Chemosphere* 93(10):2650–2655. <https://doi.org/10.1016/j.chemosphere.2013.09.089>
11. Toropova AP, Toropov AA, Benfenati E, Puzyn T, Leszczynska D, Leszczynski J (2014) *Ecotoxicol Environ Saf* 108:203–209. <https://doi.org/10.1016/j.ecoenv.2014.07.005>
12. Toropov AA, Toropova AP, Benfenati E, Gini G, Puzyn T, Leszczynska D, Leszczynski J (2012) *Chemosphere* 89(9):1098–1102. <https://doi.org/10.1016/j.chemosphere.2012.05.077>
13. Toropova AP, Toropov AA, Beeg M, Gobbi M, Salmona M (2017) *Curr Drug Discov Technol* 14(4):229–243. <https://doi.org/10.2174/1570163814666170525114128>
14. Toropova MA, Veselinović AM, Veselinović JB, Stojanović DB, Toropov AA (2015) *Comput Biol Chem* 59:126–130. <https://doi.org/10.1016/j.compbiolchem.2015.09.009>
15. Duchowicz PR, Fioressi SE, Bacelo DE, Saavedra LM, Toropova AP, Toropov AA (2015) *Chemom Intell Lab Syst* 140:86–91. <https://doi.org/10.1016/j.chemolab.2014.11.008>
16. Toropova AP, Toropov AA, Kudyshkin VO, Leszczynska D, Leszczynski J (2014) *J Math Chem* 52(5):1171–1181. <https://doi.org/10.1007/s10910-014-0323-3>
17. Wu W, Zhang R, Peng S, Li X, Zhang L (2016) *Chemom Intell Lab Syst* 157:7–15. <https://doi.org/10.1016/j.chemolab.2016.06.011>
18. Toropov AA, Toropova AP, Martyanov SE, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2011) *Chemom Intell Lab Syst* 109(1):94–100. <https://doi.org/10.1016/j.chemolab.2011.07.008>
19. Roy K, Kar S, Das RN (2015) *Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment*. Academic Press. <https://doi.org/10.1016/C2014-0-00286-9>
20. Toropova AP, Toropov AA, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2011) *J Comput Chem* 32(12):2727–2733. <https://doi.org/10.1002/jcc.21848>
21. Toropova AP, Toropov AA (2022) *Sci Total Environ* 823:153747. <https://doi.org/10.1016/j.scitotenv.2022.153747>
22. Toropova AP, Toropov AA, Leszczynska D, Leszczynski J (2021) *Comput Biol Med* 136:104720. <https://doi.org/10.1016/j.compbiomed.2021.104720>
23. Toropov AA, Toropova AP (2021) *Sci Total Environ* 772:145532. <https://doi.org/10.1016/j.scitotenv.2021.145532>
24. Toropova AP, Toropov AA, Rallo R, Leszczynska D, Leszczynski J (2016) *Int J Environ Res* 10(1):59–64. <https://doi.org/10.22059/IJER.2016.56888>
25. Toropova AP, Toropov AA, Fjodorova N (2022) *NanoImpact* 28:100427. <https://doi.org/10.1016/j.impact.2022.100427>
26. Toropova AP, Toropov AA, Leszczynski J, Sizochenko N (2021) *Environ Toxicol Pharmacol* 86:103665. <https://doi.org/10.1016/j.etap.2021.103665>
27. Toropova AP, Toropov AA (2021) *Int J Environ Res* 15(4):709–722. <https://doi.org/10.1007/s41742-021-00346-w>
28. Toropova AP, Toropov AA, Benfenati E (2019) *Fuller Nanotub Carbon Nanostruct* 27(10):816–821. <https://doi.org/10.1080/1536383x.2019.1649659>
29. Toropova AP, Toropov AA, Benfenati E, Castiglioni S, Bagnati R, Passoni A, Zuccato E, Fanelli R (2018) *Process Saf Environ Prot* 118:227–233. <https://doi.org/10.1016/j.psep.2018.07.003>
30. Lotfi S, Ahmadi S, Kumar P (2022) *RSC Adv* 12(38):24988–24997. <https://doi.org/10.1039/D2RA03936B>
31. Toropova AP, Toropov AA, Roncaglioni A, Benfenati E (2022) *SAR QSAR Environ Res* 33(8):621–630. <https://doi.org/10.1080/1062936x.2022.2104369>
32. Kumar P, Kumar A, Singh D (2022) *Environ Toxicol Pharmacol* 93:103893. <https://doi.org/10.1016/j.etap.2022.103893>

33. Toropova AP, Toropov AA, Roncaglioni A, Benfenati E (2022) <https://doi.org/10.21203/rs.3.rs-1744436/v1>
34. Ruark CD, Hack CE, Robinson PJ, Anderson PE, Gearhart JM (2013) *Arch Toxicol* 87(2):281–289. <https://doi.org/10.1007/s00204-012-0934-z>
35. Toropova AP, Toropov AA, Roncaglioni A, Benfenati E (2022) *Chem Phys Lett* 790:139354. <https://doi.org/10.1016/j.cplett.2022.139354>
36. Achary PGR, Toropova AP, Toropov AA (2019) *Int Food Res J* 122:40–46. <https://doi.org/10.1016/j.foodres.2019.03.067>
37. Toropov AA, Toropova AP, Marzo M, Benfenati E (2020) *J Mol Graph* 96:107525. <https://doi.org/10.1016/j.jmgm.2019.107525>
38. Toropova AP, Toropov AA, Carnesecchi E, Benfenati E, Dorne JL (2020) *Environ Sci Pollut Res* 27(12):13339–13347. <https://doi.org/10.1007/s11356-020-07820-6>
39. Achary PGR, Toropova AP, Toropov AA (2021) *Process Saf Prog* 40(2):e12189. <https://doi.org/10.1002/prs.12189>
40. Ahmadi S, Ketabi S, Qomi M (2022) *New J Chem* 46(18):8827–8837. <https://doi.org/10.1039/d2nj00596d>
41. Singh R, Kumar P, Devi M, Lal S, Kumar A, Sindhu J, Toropova AP, Toropov AA, Singh D (2022) *New J Chem* 46:19062–19072. <https://doi.org/10.1039/d2nj03515d>
42. Toropov AA, Toropova AP, Kudyshkin VO, Bozorov NI, Rashidova SS (2020) *Struct Chem* 31(5):1739–1743. <https://doi.org/10.1007/s11224-020-01588-8>
43. Kumar P, Kumar A, Lal S, Singh D, Lotfi S, Ahmadi S (2022) *J Mol Struct* 1265:133437. <https://doi.org/10.1016/j.molstruc.2022.133437>
44. Kumar A, Kumar P, Singh D (2022) *Chemom Intell Lab Syst* 224:104552. <https://doi.org/10.1016/j.chemolab.2022.104552>
45. Toropova AP, Toropov AA, Carnesecchi E, Benfenati E, Dorne JL (2020) *Chem Pap* 74:601–609. <https://doi.org/10.1007/s11696-019-00903-w>
46. Toropov AA, Toropova AP, Achary PGR, Raskova M, Raska I (2022) *Toxicol Mech Methods* 32(7):549–557. <https://doi.org/10.1080/15376516.2022.2053918>

Part VI
Possible Ways of Nano-QSPR/Nano-QSAR
Evolution

Chapter 14

The CORAL Software as a Tool to Develop Models for Nanomaterials' Endpoints



Alla P. Toropova and Andrey A. Toropov

Abstract This chapter discusses the evolution of the so-called quasi-SMILES. The traditional simplified molecular-input line-entry system (SMILES) is a string of characters conveying information about the structure of molecules. Quasi-SMILES is a string of characters that can convey codes reflecting the structure of molecules and the conditions for conducting chemical or biochemical experiments. Several examples demonstrate the similarity in reporting data on individual nanomaterials and data on two or more nanomaterials subjected to the same type of experiment. The possibility of gradual expansion of the scope of application of quasi-SMILES, as well as the possibility of using quasi-SMILES as input information for the CORAL software (abbreviation CORrelation And Logic) when building models of physicochemical and biochemical phenomena for nanomaterials, is shown.

Keywords Nano-QSPR · Nano-QSAR · Quasi-SMILES · Monte Carlo method · CORAL software

14.1 Introduction

In their autobiography, Sir Harry Kroto (Nobel Prize, 1996) noted, '... I had the strong gut feeling that it was so beautiful a solution that it just had to be right.' It is about fullerene structure C_{60} . Although the practical applications of C_{60} have remained limited, its discovery changed the perception of the behavior of carbon and paved the way for work on carbon nanotubes and graphene. The existence and formation of C_{60} molecules in outer space were detected and confirmed; hence, the astronomical role of C_{60} is established [1–3]. Analysis of cave paintings suggests that people of ancient civilizations used nanomaterials, such as graphene, without knowing it [4].

A. P. Toropova · A. A. Toropov (✉)
Laboratory of Environmental Chemistry and Toxicology, Department of Environmental Health Science, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156 Milano, Italy
e-mail: andrey.toropov@marionegri.it

Luster ceramic decoration was revealed by analytical electron microscopy to have been the first nanostructured film made by man. This is a real technological discovery because nanocrystal films have been produced empirically since medieval times [5].

Currently, reports on nanomaterials are more common than on any other materials. There is both bad and good news. Nanomaterials are able to rapidly impact many areas of daily life, such as food, cosmetics, pharmaceuticals, electronics, building materials, medical materials, and so on. The question arises about the safety of using these still new but no longer exotic materials. Some life-threatening and health-threatening effects can be detected quickly. However, many other deleterious effects can only be detected using long-term observations or even multi-generational data.

For example, mutagenicity and/or carcinogenicity is harm that can only be detected by comparing the health of several generations. Thus, in the 'ocean of nanotechnologies,' there may be dangerous pitfalls. The dangers of some nanomaterials are currently established. It can be expected that in the near future, the list of dangerous effects of nanomaterials will expand. Nanomaterials are characterized by a strange and 'uncomfortable' molecular architecture. Traditional methods for predicting physicochemical and biological behavior are often not suitable for nanomaterials since the essence of traditional methods is to use molecular structure data to compare and further predict the behavior of substances. The behavior of traditional (small) molecules is mainly determined by the presence of various chemical elements and the configuration of covalent bonds between them. Other features resulting from the impacts of large clusters of chemical elements are likely to determine the behavior of nanomaterials. Thus, approaches aimed at predicting the behavior of nanomaterials require a new presentation of the relevant experimental data.

Analogies are practically a necessary part of research work. The transition from the study of traditional 'non-nanosubstances' to the study of nanomaterials is analogous to the transition from considering the economic state of villages to considering the economic state of cities [6] or the transition from looking at calculators to looking at computers. The village may be loosely connected to other parts of the planet. The city must be connected to other cities. The economic status of the village is determined by the ratio of men, women, and children: a small number of workers, as a rule, leads to a decrease in the economic potential of the village. In the case of a city, these criteria are not informative. However, it is possible to define specific indicators of the economic potential of the city (not informative for the countryside), for example, the number of stations and airports. If we continue this 'village-city' analogy, then we can state that in the case of traditional substances, the basis for predicting the physicochemical and/or biological potential of a substance ('village') is a comparison of the molecular structure of this substance with the molecular structure of other similar substances (analogy proportion of men, women, and children in villages), while in the case of nanomaterials, other characteristics must also be compared. The conditions of synthesis and the conditions of the impact of nanomaterials on biological targets (cells, membranes, organs, animals, humans) are informative characteristics analogous to the 'number of stations and airports' for cities.

Thus, developing models of nanomaterials' physicochemical and biochemical behavior is a real and important task of modern natural sciences. Previously, the

solution to this problem was expected in the form of a paradigm like QSPR/QSAR (quantitative structure–property/activity relationships) [7]. However, apparently, such a solution will not have complete similarity with QSPR/QSAR, although some analogies are quite possible.

The Organization for Economic Co-operation and Development (OECD) has numerous goals concerning the development of international cooperation in the field of economy and ecology, as well as in the field of natural and human sciences. The appearance of new categories of nanomaterials implies radical modifications of methods of computational modeling physicochemical and biochemical endpoints desired. For such a task, the traditional QSPR/QSAR approaches need a radical transformation.

Modern society seems to increase its rate of risk production constantly (i.e., industrial and agricultural pollutions, destroy of ecosystems via technological disasters and others), and this is not only due to the increased production of advanced technology [8]. There are four components of any real risk assessment: identification, risk analysis, risk impact, and economic aspect of the development of the corresponding legislation documents. According to OECD, dissolution rate and dispersion stability in the environment are important parameters for nanomaterials, i.e., these parameters are the main drivers in the environmental fate of nanomaterials and nanomaterials (bio)availability [9]. Therefore, the development of models for other endpoints related to nanomaterials is a practical task that also is significant and urgent [7, 10, 11].

The problem of assessing the risk of using nanomaterials in the environmental aspect intersects with the problem of the correct, efficient, and safe use of nanotechnologies in medicine [12, 13]. Factually, medicine involves fullerenes [14–16], single carbon nanotubes [17, 18], multiwall carbon nanotubes [19, 20], nano-oxides [21, 22], and quantum dots [23, 24].

Nanomaterials are widely used in cosmetics [25–31]. However, the lists of nanomaterials for medicine and cosmetics are pretty different. Nano-oxides are mainly used for cosmetics [27, 28] and, to a lesser extent, also fullerenes [30]. Concerns over health risks have limited the further incorporation of nanomaterials in cosmetics. Since the cosmetic industry may use new nanomaterials in the future, a detailed characterization and risk assessment are needed to fulfill the standard safety requirements. To solve the above task, undouble the fast methods of risk assessment using computational approaches, which are currently being developed [31]. It is to be noted that the stream of nanotechnology applications involves not only medicine and cosmetics but also electronics [23, 24] and even the design of nanorobots [32].

The term nano-QSAR appears for the first time in work of Puzyn et al. [33]. Thus, the efforts of researchers aimed at the development of nano-QSAR started less than fifteen years ago. Pretty soon, it became clear that some qualitative changes in the QSAR paradigm were needed for the cases of nanomaterials. It became obvious that a new approach was needed to re-define databases that were suitable for classical QSAR but were not suitable for the case of QSAR for nanomaterials. To solve this problem, an ISA-TAB-nanoparadigm (investigation–study–assay) was proposed [34]. The approach is based on the representation of nano-data in a

particular format, ‘investigation–study–assay’ [34]. The term ‘nano-informatics’ was suggested perhaps to the same end [35].

Another critical point in the search for approaches for modeling the physico-chemical properties and biological activity of nanomaterials is the search for ways to consolidate potential consumers of nano-models online, through special websites. For instance, to make the nano-model (the model is based on the k-Nearest Neighbors (kNN) algorithm) available to interested users, the model was made available via the Internet (Enalos In Silico Nanoplatform [36]). Quantum mechanical descriptors as a basis for nano-QSAR have been successfully used to model nano-oxides toxicity [37]. The k-nearest neighbors (kNN-based regression) and support vector machine (SVM) were applied to build up a good model for PaCa2 cell line uptake on 109 nanoparticles [38].

Having a group of records related to the influence of nanomaterials upon the biological targets under different conditions, one can select conditions of three categories of their impacts: (i) conditions that are promoters of increased impact; (ii) conditions that are promoters of decreased impact; and (iii) conditions that do not influence the impact of nanomaterials. The CORAL software (<http://www.insilico.eu/coral>) gives the possibility to automatically carry out the analysis of the records related to nanomaterials, mentioned above. Moreover, it is possible to integrate separated recommendations into a united system of estimation for a large group of different nanomaterials in the future.

Numerous disputes about the expediency of constructing quantitative structure–property/activity relationships (QSPRs/QSARs) have not yet led to a denial of the main issue—that such studies are useful both in practical and theoretical terms. At the beginning of its development, the theory and practice of the QSPR/QSAR research were criticized for the lack of transparency in the interpretation of models [38–40]. However, later, the questionable reliability and reproduction of the models became the main point of criticism [41–44].

Nevertheless, QSPR/QSAR method has gradually become a generally accepted tool for constructing models of physicochemical properties and biological activity for organic [44–47], inorganic [48], organometallic [49] compounds, and polymers [50]. The listed categories of substances are characterized by unambiguous molecular structure, which, in fact, is the basis for constructing QSPR/QSAR models.

However, in the case of models related to nanomaterials [51], the representation of an exclusively molecular structure or even in conjunction with data from molecular mechanics and quantum chemistry calculations [52] becomes insufficient for the development of new perspective ways of building up models for phenomena in biology, and medicinal nanotechnologies [53]. New technologies used in the agriculture and food industry, e.g., nano-pesticides, force a revision of the QSAR and the quantum mechanics as well all other descriptors suitability, owing to the high diversity of dangerous effects which can be observed in the case of using nano-pesticides [53].

Significant difficulties arise from the fact that a small change in the production of such pesticides can lead to a significant change in their impact on biosystems, ecosystems or even economic systems. In other words, an agricultural process based

on nanotechnology may become both environmentally and economically unsustainable, as the unclear impact on yields will be a danger to ecosystems and human health. Another important point is the increase in the diversity of nanomaterials. For example, in recent years, nano-cellulose has attracted increasing attention from researchers and industry as an alternative to traditional cellulose [54]. The same situation occurs for quantum dots, which are becoming more and more widespread in research and industry [55]. Obviously, a wide variety of nanomaterials inevitably leads to a greater likelihood of unexpected and often unpleasant or, moreover, dangerous effects.

A sufficiently detailed analysis of a large number of various scenarios is economically complex. Therefore, for example, the development of reliable models of the physicochemical and biochemical behavior of quantum dots associated with nanomaterials is an evident and important task. It should be taken into account that these models should accurately reflect the experimental conditions. The choice of a list of experimental conditions that should be available for building a model is also a non-trivial and important problem.

Unfortunately, the solution to the above problem likely should be selective and tuning for each specific experiment. Thus, consideration of a new nanomaterial category can imply radical modifications of QSPR and QSAR conception.

QSPR/QSAR studies obey special rules defined by different international organizations (e.g., the above OECD). These rules supply particular standards to make the corresponding models and provide software reliable enough for practical use [56, 57]. However, it should be noted the above standards are not dogmas. Moreover, these standards will develop and change rapidly. Since the theory and practice of nanomaterials manufacturing are innovative, these standards will change and improve according to new experimental data on the abilities and dangerousness of nanomaterials.

14.2 Theory and Practices of QSPR/QSAR

Any QSPR/QSAR model implies a way of estimating of value of a parameter of interest to a substance y via a mathematical function

$$y = F(\text{all available influences on the system}). \quad (14.1)$$

In the classical QSPR/QSAR, the equation defined as

$$y = F(\text{all available descriptors}). \quad (14.2)$$

The predictive capability of the QSAR models is established by performing an external validation, information indices, topological indices, quantum mechanics descriptors, molecular operating environment indices, and just physicochemical parameters of substances under investigation. It has been shown that those measures are appropriate tools for selecting the model calculated with Eq. 14.2.

The advantage of the one-variable model is their reliability. Often the one-variable model is characterized by a modest (but not poor) statistical quality for both the training and test sets. Multiple linear regression analysis (MLRA) can be utilized to obtain a model that is developed using a group of several descriptors. The model obtained by MLRA is often characterized by good (perfect) statistical quality for the training set, but this model can be a poor one for the external validation set.

According to many authors, a rational split data into training and validation sets gives better statistical results for the validation sets than models based on random splits. However, the experiment confirms that often there are some distributions into the training and validation sets successful for one approach, which is unsuccessful for another method.

All the above-mentioned circumstances and instructions become precise and reliable basis for building up models for the physicochemical and biochemical behavior of diverse substances. Nevertheless, in the science space, nanomaterials have become a new, unexpected scientific targets. The theory and practices of the QSPR/QSAR require new tools to analyze these new substances. There are, however, principal barriers that make corresponding efforts ineffective.

First, molecules of nanomaterials are large; more exactly, nanomaterials' molecules are much larger than molecules of most traditional substances (which are not nanomaterials). Second, the difference between the physicochemical or biochemical behavior of the two nanomaterials is caused rather by an influence of the medium and not by intramolecular interactions.

On the one hand, according to Bertrand Russell, 'All exact science is dominated by the idea of approximation'; on the other hand, 'all models are wrong but some are useful' [58].

Gradually, the target of the QSPR/QSAR research shifted from the selection of a perfect molecular structure to the harmonization of all available often-eclectic circumstances. For example, a drug should not be toxic. Cosmetics should not be bio-accumulative. Plastic should be biodegradable.

Applying QSPR/QSAR for regulatory aims is an attractive idea. Still, this idea is hardly realized since, for regulatory purposes, the experiment is the only way to get the necessary numerical data and technical information. Computational experiments aimed at estimating toxicity are surrogates of real experiments on toxicity assessment. No one could declare data on toxicity to be reliable if the data is provided from mathematical methods, without verification by corresponding experiments. Economic and legal evaluation of a new substance is available only based on a real experiment.

The different methodologies aimed to solve the above tasks in the ecologic risk assessment hardly can be systematized. In other words, the current QSPR/QSAR as well as the QSPR/QSAR in the future become a mathematical function of eclectic data, not solely a mathematical function of the molecular structure. Figure 14.1 illustrates the trend.

Nevertheless, the various phenomena observed in computer experiments aimed at building QSPR/QSAR models are sometimes very similar to those observed in traditional science experiments, performed without computers. For example, the dependence of the numbers of poor predictions and percentage of poor predictions

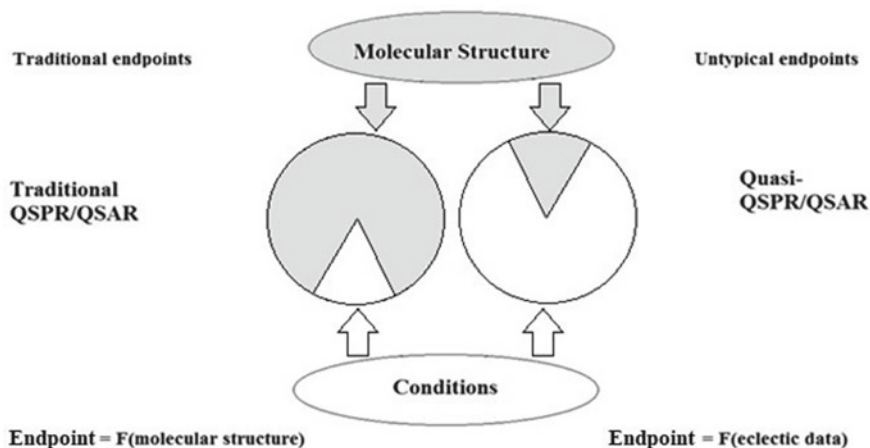


Fig. 14.1 Essence of QSPR/QSAR models

for the data validation set in a group of models observed for different splits into the training and validation sets is similar to the dependence of conductivity and thickness in molybdenum disulfide MoS_2 nanoflakes [59].

14.3 SMILES and Nanomaterials

Paradoxically, the practical development of QSPR/QSAR models related to nanomaterials began without any data on the molecular structure of nanomaterials [60–62]. These were models aimed at predicting the solubility of fullerene (C_{60}) in various organic solvents. Thus, although the solubility of the nanomaterial was modeled, its molecular structure was not used to develop this model. The only fact of the presence of fullerene in the solution was considered in this research.

Nevertheless, soon after initial works the comparison of the regularities of biological activity along different sequences of fullerene derivatives (C_{60}) with considering molecular features of fullerene derivatives [63–65] began. Similar research was carried out for single-wall carbon nanotubes (SWCNTs) [66, 67] and multi-walls carbon nanotubes (MWCNTs) [68].

The unique qualities of SWCNTs, MWCNTs, and fullerenes C_{60} and C_{70} required an actual revision of the traditional concept of QSPR/QSAR as models based on molecular 2D and 3D descriptors obtained from molecular mechanics and quantum mechanical calculations. Instead, descriptors partially reflecting 'nano-nature' of substances were developed and tested. Such approach subsequently became almost independent of the molecular structure, concentrating on the experimental conditions [69–71]. Nevertheless, it is to be noted that some QSPR/QSAR research is based

solely on the molecular structure of fullerene derivatives without considering the conditions of an experiment carried out [61, 64, 72–76].

Fullerene derivatives have been studied longer than most other nanomaterials. This led to a considerable flow of work devoted to these substances, which were considered exotic for a long time. At present, fullerene derivatives have found some applications. These applications do the QSPR/QSAR analysis of fullerene derivatives actual. The first attempts at such an analysis were carried out using a simplified molecular-input line-entry system (SMILES) for anti-HIV activity [72], the solubility of C₆₀ in organic compounds [61], and mutagenicity of fullerene derivatives [73]. Further computation experiments dedicated to building up predictive models were aimed to extend the targets list, namely to the mutagenicity of SWCNTs and fullerene C₆₀ [76], and united models for mutagenicity of fullerenes C₆₀ and C₇₀ [64].

SMILES is the representation of the molecular structure by a sequence of special symbols that encode different molecular features such as atoms, bonds, presence/absence of various rings [77]. Briefly, the SMILES is a line where chemical elements represented by corresponding symbols (e.g., 'C' = carbon; 'Br' = bromine, etc.); double covalent bonds indicated by '=', triple covalent bonds indicated by '#,' as well there are some other special codes for combination of rings (e.g., digits 1–9, and %10, %11, etc.); finally, some 3D features also taken into account (e.g., @, or @@). It can be said that, at present, there is some implicit competition between SMILES and graphs in the development of models of various physicochemical and biological parameters for various molecular systems. In some cases, it is preferable to use molecular graphs. In other cases, SMILES is more convenient. At the same time, an important circumstance is that these representations are far from identical. However, they are aimed at solving the same task, namely representing molecular structures in databases, providing users with the ability to quickly identify and compare all kinds of molecular features [78–86].

14.4 Quasi-SMILES and Nanomaterials

When developing any program, one should know the answers to several questions. Table 14.1 contains a collection of such questions as well as some typical responses to these.

Everything listed in Table 14.1 refers to SMILES as a tool for solving practical problems in mathematical and computational chemistry. Unfortunately, during the development of quasi-SMILES, logistics were not planned at all. It was assumed that quasi-SMILES is a tool for creating models according to the paradigm expressed by Eq. 14.2, i.e., quasi-SMILES was aimed to include maximum information to develop a model.

Just as in traditional molecules, the presence of various fragments down to individual atoms and bonds affects molecules' ability to be solvents, poisons, or something else. Additionally, the presence or absence of light, the concentration

Table 14.1 Logistics of development QSPR/QSAR aimed software

Questions	Responses
What is the model that is expected to build up?	The model makes it possible, having some list of the features of the phenomenon, to predict how the situation will change if the values of the mentioned features are changed or the list is changed (expanded or shortened)
For whom is the model?	The model users will be those interested in the opportunity to influence the phenomenon under consideration (experimentations); those who develop similar models; those who plan to be an experimentations or developers of such models
How to provide the model to potential consumers?	It is obvious that, first, one should be informed about what this program can accomplish for the potential user, and second, information should be available on how to use it
Does the software developer need user feedback?	If the development of the program is planned, then feedback is needed
How do establish feedback with consumers?	The only option is dialogue. Dialogue is actually possible only if the user wants it

of impurities, and the nature of porosity can affect the ability of nanomaterials (physicochemical, biochemical, and others).

Thus, at the very beginning, quasi-SMILES gave models for the behavior of nanomaterials depending on the experimental conditions, while the molecular architecture of nanomaterials was not involved in the development of the model at all [69, 87]. Table 14.2 contains an example of the list of experimental conditions used as a basis to build up such models.

However, later, quasi-SMILES were improved by including codes indicating various nanomaterials [69, 70, 88]. It can be interpreted as 'fullerene acts here' or 'multi-walled carbon nanotubes act here.' The collection of such experimental conditions is represented in Table 14.3. Table 14.4 contains an example of codes for

Table 14.2 List of attributes of fullerene C₆₀ nanoparticles exposure and their codes which are used for the construction of quasi-SMILES

Experimental conditions	Codes for quasi-SMILES
The presence or absence of lighting	The code '0' means absence of lighting The code '1' means presence of lighting
Mix S9	The code '+' means 'with mix S9' The code '-' means 'without mix S9'
Dose	The code 'A' means the dose 50 g/plate The code 'B' means the dose 100 g/plate The code 'C' means the dose 200 g/plate The code 'D' means the dose 400 g/plate The code 'E' means the dose 1000 g/plate

Table 14.3 Brunauer–Emmett–Teller (BET) surface area analysis: an example of experimental conditions on bacterial reverse mutation tests on multi-walled carbon nanotubes of two types [88]

MWCNT ^a Diameter, nm/BET	Surface area, m ² /g	Concentration, μg/plate	S9 microsomal fraction	The average number of revertant colonies/plate, TA100
44	69	0.78	Without mix S9	120
44	69	1.56	Without mix S9	109
44	69	3.13	Without mix S9	119
44	69	6.25	Without mix S9	116
44	69	12.5	Without mix S9	114
44	69	25.0	Without mix S9	109
44	69	50.0	Without mix S9	114
44	69	100.0	Without mix S9	117
44	69	0.78	With mix S9	105
44	69	1.56	With mix S9	115
44	69	3.13	With mix S9	114
44	69	6.25	With mix S9	127
44	69	12.5	With mix S9	133
44	69	25.0	With mix S9	120
44	69	50.0	With mix S9	125
44	69	100.0	With mix S9	128
70	23	0.78	Without mix S9	111
70	23	3.13	Without mix S9	118
70	23	6.25	Without mix S9	122
70	23	12.5	Without mix S9	123
70	23	25.0	Without mix S9	118
70	23	50.0	Without mix S9	121
70	23	100.0	Without mix S9	121
70	23	0.78	With mix S9	126
70	23	3.13	With mix S9	114
70	23	6.25	With mix S9	135
70	23	12.5	With mix S9	124
70	23	25.0	With mix S9	124
70	23	50.0	With mix S9	108
70	23	100.0	With mix S9	134

^a N-MWCNT (diameter = 44 and surface area 69); MWNT-7 (diameter = 70 and surface area 23)

Table 14.4 List of codes used to construct quasi-SMILES reflecting the situation where two kinds of multi-walled carbon nanotubes act in similar experimental conditions

Experimental conditions	Codes to construct quasi-SMILES
Test substance	The code '1' means presence of N-MWCNT The code '2' means presence of MWNT-7
Mix S9	The code '+' means 'with mix S9' The code '-' means 'without mix S9'
Concentration	The code 'A' means the dose 0.78 $\mu\text{g}/\text{plate}$ The code 'B' means the dose 1.56 $\mu\text{g}/\text{plate}$ The code 'C' means the dose 3.13 $\mu\text{g}/\text{plate}$ The code 'D' means the dose 6.25 $\mu\text{g}/\text{plate}$ The code 'E' means the dose 12.5 $\mu\text{g}/\text{plate}$ The code 'F' means the dose 25.0 $\mu\text{g}/\text{plate}$ The code 'G' means the dose 50.0 $\mu\text{g}/\text{plate}$ The code 'H' means the dose 100.0 $\mu\text{g}/\text{plate}$

constructing quasi-SMILES reflecting the situation where two kinds of multi-walled carbon nanotubes act under similar experimental conditions.

Using codes (Table 14.4), one can obtain a predictive system represented by Table 14.5. One can see the result of three different distributions of data in the training set (T), calibration set (C), and validation set (V).

In fact, traditional SMILES uses a significant portion of the available characters. Under such circumstances, certain compromises had to be found to search for a letter (symbol) basis for quasi-SMILES constructions. Particular agreed-upon combinations such as $A_1, A_2, \dots, A_9, B_1, B_2, \dots, B_9, \dots$ were used to discretize various scales.

The examples in Tables 14.2, 14.3, 14.4, 14.5, 14.6, and 14.7 showed similar situations when 10–15 additional special characters were enough to develop quasi-SMILES and corresponding models. In principle, the collection of such models can be expanded with new analogous models of the physicochemical properties or biological activity of nanomaterials [89, 90], peptides [91], or membranes [92]. However, this approach is not comfortable for users (limited number of special characters, weak mnemonics, etc.). To increase comfort, there was an attempt to involve special groups of symbols borrowed directly from the classic SMILES. These groups of symbols aimed to represent in SMILES information about the presence of more than ten rings [77], e.g., %11, %12, etc. (molecule contains eleven, twelve, or more rings, respectively). Figure 14.2 includes some examples of discretion of a parameter (experimental condition) to involve in quasi-SMILES.

The discretion representation for a parameter X is calculated using the formula Eq. 14.3 [93–95].

$$\text{Discret}(X) = \frac{X_{\min} + X_k}{X_{\min} + X_{\max}} \quad (14.3)$$

Table 14.5 Three distributions of available experimental data into the training (T), calibration (C), and validation (V) sets; three-symbols quasi-SMILES representing genotoxicity by multi-walled carbon nanotubes, experimental and predicted average TA100 values (the number of revertant colonies/plate)

ID	Split			Quasi-SMILES	TA100		
	1	2	3		Experiment	Average prediction	Dispersion
01	C	V	T	1-A	120	111.98	± 9.22
02	T	T	T	1-B	109	108.19	± 1.67
03	C	T	T	1-C	119	112.96	± 7.62
04	V	T	C	1-D	116	119.16	± 2.08
05	T	C	T	1-E	114	118.34	± 5.06
06	T	V	C	1-F	109	111.94	± 6.34
07	T	C	T	1-G	114	116.98	± 4.04
08	V	T	V	1-H	117	119.25	± 2.01
09	T	T	T	1+A	105	116.03	± 0.53
10	V	T	V	1+B	115	119.86	± 5.16
11	V	V	T	1+C	114	118.32	± 3.39
12	V	V	C	1+D	127	135.19	± 2.36
13	T	T	T	1+E	133	128.83	± 0.95
14	C	C	V	1+F	120	123.06	± 0.42
15	T	T	T	1+G	125	116.93	± 0.54
16	C	T	V	1+H	128	132.60	± 2.04
17	T	C	T	2-A	111	115.33	± 5.36
18	T	V	C	2-C	118	116.31	± 4.10
19	T	T	T	2-D	122	122.52	± 4.46
20	T	C	T	2-E	123	121.68	± 2.85
21	T	T	C	2-F	118	115.29	± 5.54
22	T	T	T	2-G	121	120.33	± 1.84
23	T	V	T	2-H	121	122.60	± 1.97
24	T	T	T	2+A	126	115.11	± 0.32
25	T	T	T	2+C	114	117.40	± 3.43
26	T	T	T	2+D	135	134.26	± 3.14
27	T	V	T	2+E	124	127.90	± 0.12
28	C	T	V	2+F	124	122.14	± 1.23
29	T	T	T	2+G	108	116.01	± 0.30
30	T	T	T	2+H	134	131.68	± 1.83

Table 14.6 Features of action of nanomaterials (fullerene and MWCNT) and their codes

ID	Feature	Code for the feature
I	Fullerene MWCNT	The code 'X' means presence of fullerene The code 'Z' means presence of MWCNT
II	Dark or irradiation	The code '0' means presence of dark The code '1' means presence of irradiation
III	Preincubation	The code 'N' means absence of preincubation The code 'Y' means presence of preincubation
IV	Mix S9	The code '+' means 'with mix S9' The code '-' means 'without mix S9'
V	Dose	<i>Fullerene</i> The code 'A' means the dose 50 g/plate The code 'B' means the dose 100 g/plate The code 'C' means the dose 200 g/plate The code 'D' means the dose 400 g/plate The code 'E' means the dose 1000 g/plate <i>MWCNT</i> The code 'F' means the dose 0 μ g/plate The code 'G' means the dose 50 μ g/plate The code 'H' means the dose 158 μ g/plate The code 'I' means the dose 500 μ g/plate The code 'J' means the dose 1581 μ g/plate The code 'K' means the dose 5000 μ g/plate

Table 14.7 Construction of quasi-SMILES for the study of fullerene and MWCNT under the same experimental conditions

No.	I	II	III	IV	V	Quasi-SMILES	pTA100
1	X	0		+	A	X0+A	- 2.1640
2	X	0		+	B	X0+B	- 2.1490
3	X	0		+	C	X0+C	- 2.2010
4	X	0		+	D	X0+D	- 2.2040
5	X	0		+	E	X0+E	- 2.2480
6	X	0		-	A	X0-A	- 2.1550
7	X	0		-	B	X0-B	- 2.1430
8	X	0		-	C	X0-C	- 2.2280
9	X	0		-	D	X0-D	- 2.2250
10	X	0		-	E	X0-E	- 2.1820
11	X	1		+	A	X1+A	- 2.1110
12	X	1		+	B	X1+B	- 2.1170
13	X	1		+	C	X1+C	- 2.1400

(continued)

Table 14.7 (continued)

No.	I	II	III	IV	V	Quasi-SMILES	pTA100
14	X	1		+	D	X1+D	- 2.1370
15	X	1		+	E	X1+E	- 2.2040
16	X	1		-	A	X1-A	- 2.1340
17	X	1		-	B	X1-B	- 2.1340
18	X	1		-	C	X1-C	- 2.1400
19	X	1		-	D	X1-D	- 2.2150
20	X	1		-	E	X1-E	- 2.2360
21	Z		N	-	F	ZN-F	- 2.0830
22	Z		N	-	G	ZN-G	- 2.1140
23	Z		N	-	H	ZN-H	- 2.0830
24	Z		N	-	I	ZN-I	- 2.0930
25	Z		N	-	J	ZN-J	- 2.0450
26	Z		N	-	K	ZN-K	- 1.9730
27	Z		Y	-	F	ZY-F	- 2.1210
28	Z		Y	-	G	ZY-G	- 2.1240
29	Z		Y	-	H	ZY-H	- 2.1000
30	Z		Y	-	I	ZY-I	- 2.1140
31	Z		Y	-	J	ZY-J	- 2.1070
32	Z		Y	-	K	ZY-K	- 2.0900
33	Z		N	+	F	ZN+F	- 2.1340
34	Z		N	+	G	ZN+G	- 2.1550
35	Z		N	+	H	ZN+H	- 2.1340
36	Z		N	+	I	ZN+I	- 2.1040
37	Z		N	+	J	ZN+J	- 2.0680
38	Z		N	+	K	ZN+K	- 2.0570
39	Z		Y	+	F	ZY+F	- 2.2740
40	Z		Y	+	G	ZY+G	- 2.2740
41	Z		Y	+	H	ZY+H	- 2.2790
42	Z		Y	+	I	ZY+I	- 2.2600
43	Z		Y	+	J	ZY+J	- 2.2430
44	Z		Y	+	K	ZY+K	- 2.2380

The anatomy of the model is based on four-symbol quasi-SMILES [70]

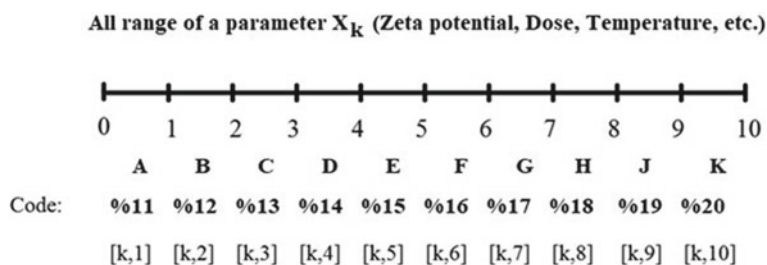


Fig. 14.2 Different versions of discretion of a parameter to involve in quasi-SMILES

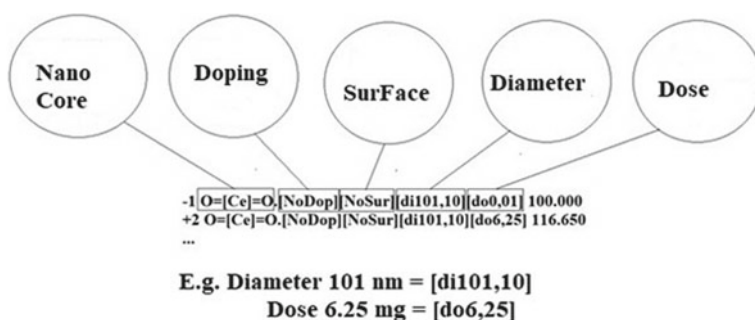


Fig. 14.3 The scheme of building up quasi-SMILES

Another way to construct quasi-SMILES codes has been suggested recently [96]. Such approach gives the possibility to encode experimental conditions by special codes in squared brackets. Figure 14.3 shows the general scheme of application of this approach.

14.5 Optimal SMILES-Based Descriptor

The practical realization of the approach expressed by Eq. 14.2 is the so-called optimal descriptor. The optimal descriptor is calculated with special coefficients named the correlation weights. These coefficients are calculated by the Monte Carlo method. The above-mentioned calculations can be implemented by the free-downloading CORAL software available on the Internet (<http://www.insilico.eu/coral>).

The optimal quasi-SMILES-based descriptor $DCW(T, N)$ is applied for a predictive model of the endpoint via the equation:

$$\text{Endpoint} = C_0 + C_1 \times DCW(T, N) \quad (14.4)$$

$$\text{DCW}(T, N) = \sum \text{CW}(S_k). \quad (14.5)$$

T is a threshold, that is, an integer separating quasi-SMILES attributes into two classes, 'rare' and 'non-rare.' Only 'non-rare' quasi-SMILES attributes are used to build the model. N is the number of epochs of the Monte Carlo method optimization of correlation weights. S_k is a quasi-SMILES-atom, i.e., a single character of a quasi-SMILES string (e.g., '=', 'O') or a group of characters that cannot be treated in isolation (e.g., 'Cu,' '%11'). $\text{CW}(S_k)$ are the correlation weights of the above quasi-SMILES attributes.

14.6 The Monte Carlo Optimization

Equation 14.5 requires the numerical data on the above correlation weights. Monte Carlo optimization is a tool to calculate those correlation weights. Two different target functions for the Monte Carlo optimization are applied:

$$\text{TF} = r_{\text{AT}} + r_{\text{PT}} - |r_{\text{AT}} - r_{\text{PT}}| \times 0.1 \quad (14.6)$$

The r_{AT} and r_{PT} are correlation coefficients between the observed and predicted endpoints for the active and passive training sets, respectively. It is to be noted that the CORAL software provides some additional information on the target function and the possibility to modify and use different versions of the above target function.

Table 14.8 shows examples of applications of the optimal quasi-SMILES-based descriptors to build up models for endpoints for nanomaterials. One can see that the approach gives the significant quality of models.

Table 14.8 Statistical characteristics of nano-QSPR/QSAR models built up using the optimal quasi-SMILES-based descriptors calculated by the Monte Carlo method

Training set		Validation set		References
N	R^2	N	R^2	
25	0.55	25	0.62	[96]
66	0.85	26	0.89	[97]
–	0.70	–	0.65	[98]
–	0.99	–	0.97	[99]
17	0.97	17	0.82	[100]

14.7 Conclusions

The application of quasi-SMILES provides the possibility to involve experimental conditions as components for the calculation of a model. The quasi-SMILES has several steps of their evolution, and the evolution can be further continued. Quasi-SMILES is an approach to building models of new quality: The descriptor becomes a mathematical function of structure and experimental conditions or even a mathematical function of experimental conditions together with arbitrary circumstances that can impact the experiment results. In other words, the quasi-SMILES technique can be the source of a new way to address both theoretical and practical nanochemistry and nanobiology. In principle, quasi-SMILES can become a language of communication between experimenters and developers of the corresponding models of properties and biological activity for nanomaterials, peptides, membranes, and maybe other objects and phenomena. The CORAL software can be used as an interface for the assessment of different hypotheses (models) suggested by experimentalists analyzing situations related to nanomaterials and maybe other complex physicochemical and biochemical phenomena.

Acknowledgements This work was supported by ONTOX, grant agreement 963845 of the European Commission under the Horizon 2020 research and innovation framework program.

References

1. García-Hernández DA, Iglesias-Groth S, Acosta-Pulido JA, Manchado A, García-Lario P, Stanghellini L, Villaver E, Shaw RA, Cataldo F (2011) *Astrophys J Lett* 737(2):L30. <https://doi.org/10.1088/2041-8205/737/2/L30>
2. Iglesias-Groth S, Cataldo F, Manchado A (2011) *Mon Not R Astron Soc* 413(1):213–222. <https://doi.org/10.1111/j.1365-2966.2011.18124.x>
3. Cami J, Bernard-Salas J, Peeters E, Malek SE (2010) *Science* 329(5996):1180–1182. <https://doi.org/10.1126/science.1192035>
4. Barhoum A, García-Betancourt ML, Jeevanandam J, Hussien EA, Mekkawy SA, Mostafa M, Omran MM, Abdalla MS, Bechelany M (2022) *Nanomaterials* 12(2):177. <https://doi.org/10.3390/nano12020177>
5. Pérez-Arantegui J, Larrea A (2003) *TrAC Trends Anal Chem* 22(5):327–329. [https://doi.org/10.1016/S0165-9936\(03\)00502-8](https://doi.org/10.1016/S0165-9936(03)00502-8)
6. Atlas of Sciences. <https://atlasofscience.org/the-coral-software-as-spyglass-to-detect-coral-reefs-in-ocean-of-nanotechnologies/>. Accessed 29 July 2022
7. Villaverde JJ, Sevilla-Morán B, López-Goti C, Alonso-Prados JL, Sandín-España P (2018) *Sci Total Environ* 634:1530–1539. <https://doi.org/10.1016/j.scitotenv.2018.04.033s>
8. Hellström T (2009) *Technol Soc* 31(3):325–331. <https://doi.org/10.1016/j.techsoc.2009.06.002>
9. OECD (2020) Guidance document for the testing of dissolution and dispersion stability of nanomaterials and the use of the data for further environmental testing and assessment strategies, No. 318. ENV/JM/MONO(2020)9
10. Mu Y, Wu F, Zhao Q, Ji R, Qie Y, Zhou Y, Hu Y, Pang C, Hristozov D, Giesy JP, Xing B (2016) *Nanotoxicology* 10(9):1207–1214. <https://doi.org/10.1080/17435390.2016.1202352>

11. Lubinski L, Urbaszek P, Gajewicz A, Cronin MTD, Enoch SJ, Madden JC, Leszczynska D, Leszczynski J, Puzyn T (2013) SAR QSAR Environ Res 24(12):995–1008. <https://doi.org/10.1080/1062936X.2013.840679>
12. Chugh H, Sood D, Chandra I, Tomar V, Dhawan G, Chandra R (2018) Artif Cells Nanomed Biotechnol 46(sup1):1210–1220. <https://doi.org/10.1080/21691401.2018.1449118>
13. Marchesan S, Prato M (2013) ACS Med Chem Lett 4(2):147–149. <https://doi.org/10.1021/ml3003742>
14. Yamakoshi Y, Umezawa N, Ryu A, Arakane K, Miyata N, Goda Y, Masumizu T, Nagano T (2003) J Am Chem Soc 125(42):12803–12809. <https://doi.org/10.1021/ja0355574>
15. Castro E, Garcia AH, Zavala G, Echegoyen L (2017) J Mater Chem B 5(32):6523–6535. <https://doi.org/10.1039/c7tb00855d>
16. Anilkumar P, Lu F, Cao L, Luo PG, Liu J-H, Sahu S, Tackett KN, Wang Y, Sun Y-P (2011) Curr Med Chem 18(14):2045–2059. <https://doi.org/10.2174/092986711795656225>
17. Sacchetti C, Motamedchaboki K, Magrini A, Palmieri G, Mattei M, Bernardini S, Rosato N, Bottini N, Bottini M (2013) ACS Nano 7(3):1974–1989. <https://doi.org/10.1021/nn400409h>
18. Bhirde AA, Patel S, Sousa AA, Patel V, Molinolo AA, Ji Y, Leapman RD, Gutkind JS, Rusling JF (2010) Nanomedicine 5(10):1535–1546. <https://doi.org/10.2217/nnm.10.90>
19. Benjamin SR, Vilela RS, de Camargo HS, Guedes MIF, Fernandes KF, Colmati F (2018) Int J Electrochem Sci 13(1):563–586. <https://doi.org/10.20964/2018.01.51>
20. Wagay JA, Nayik GA, Wani SA, Mir RA, Ahmad MA, Rahman QI, Vyas D (2019) J Food Meas Charact 13(3):1805–1819. <https://doi.org/10.1007/s11694-019-00099-3>
21. Schwaminger SP, Fraga-García P, Selbach F, Hein FG, Fuß EC, Surya R, Roth H-C, Blank-Shim SA, Wagner FE, Heissler S, Berensmeier S (2017) Adsorption 23(2–3):281–292. <https://doi.org/10.1007/s10450-016-9849-y>
22. Zhong L, Yu Y, Lian H-Z, Hu X, Fu H, Chen Y-J (2017) J Nanopart Res 19(11):375. <https://doi.org/10.1007/s11051-017-4064-7>
23. Yong K-T, Law W-C, Hu R, Ye L, Liu L, Swihart MT, Prasad PN (2013) Chem Soc Rev 42(3):1236–1250. <https://doi.org/10.1039/c2cs35392j>
24. Zhang H, Yee D, Wang C (2008) Nanomedicine 3(1):83–91. <https://doi.org/10.2217/17435889.3.1.83>
25. Raj S, Jose S, Sumod US, Sabitha M (2012) J Pharm Bioallied Sci 4(3):186–193. <https://doi.org/10.4103/0975-7406.99016>
26. Nohynek GJ, Dufour EK, Roberts MS (2008) Skin Pharmacol Physiol 21(3):136–149. <https://doi.org/10.1159/000131078>
27. Lu P-J, Huang S-C, Chen Y-P, Chiueh L-C, Shih DY-C (2015) J Food Drug Anal 23(3):587–594. <https://doi.org/10.1016/j.jfda.2015.02.009>
28. Auffan M, Pedeutour M, Rose J, Masion A, Ziarelli F, Borschneck D, Chaneac C, Botta C, Chaurand P, Labille J, Bottero J-Y (2010) Environ Sci Technol 44(7):2689–2694. <https://doi.org/10.1021/es903757q>
29. Mhrranyan A, Ferraz N, Strømme M (2012) Prog Mater Sci 57(5):875–910. <https://doi.org/10.1016/j.pmatsci.2011.10.001>
30. Benn TM, Westerhoff P, Herckes P (2011) Environ Pollut 159(5):1334–1342. <https://doi.org/10.1016/j.envpol.2011.01.018>
31. Fytianos G, Rahdar A, Kyzas GZ (2020) Nanomaterials 10(5):979. <https://doi.org/10.3390/nano10050979>
32. Jiang T, Song X, Mu X, Cheang UK (2022) Sci Rep 12(1):13080. <https://doi.org/10.1038/s41598-022-17053-x>
33. Puzyn T, Leszczynska D, Leszczynski J (2009) Small 5(22):2494–2509. <https://doi.org/10.1002/sml.200900179>
34. Marchese Robinson RL, Cronin MTD, Richarz A-N, Rallo R (2015) Beilstein J Nanotechnol 6(1):1978–1999. <https://doi.org/10.3762/bjnano.6.202>
35. Panneerselvam S, Choi S (2014) Int J Mol Sci 15(5):7158–7182. <https://doi.org/10.3390/ijm15057158>

36. Melagraki G, Afantitis A (2014) RSC Adv 4(92):50713–50725. <https://doi.org/10.1039/c4ra07756c>
37. Puzyn T, Rasulev B, Gajewicz A, Hu X, Dasari TP, Michalkova A, Hwang H-M, Toropov A, Leszczynska D, Leszczynski J (2011) Nat Nanotechnol 6(3):175–178. <https://doi.org/10.1038/nnano.2011.10>
38. Fourches D, Pu D, Tassa C, Weissleder R, Shaw SY, Mumper RJ, Tropsha A (2010) ACS Nano 4(10):5703–5712. <https://doi.org/10.1021/nn1013484>
39. Doweiko AM (2008) J Comput Aided Mol Des 22(2):81–89. <https://doi.org/10.1007/s10822-007-9162-7>
40. Maggiora GM (2006) J Chem Inf Model 46(4):1535. <https://doi.org/10.1021/ci060117s>
41. Doweiko AM (2004) J Comput Aided Mol Des 18(7–9):587–596. <https://doi.org/10.1007/s10822-004-4068-0>
42. Johnson SR (2008) J Chem Inf Model 48(1):25–26. <https://doi.org/10.1021/ci700332k>
43. Dearden JC, Cronin MTD, Kaiser KLE (2009) SAR QSAR Environ Res 20(3–4):241–266. <https://doi.org/10.1080/10629360902949567>
44. Scior T, Medina-Franco JL, Do Q-T, Martínez-Mayorga K, Yunes Rojas JA, Bernard P (2009) Curr Med Chem 16(32):4297–4313. <https://doi.org/10.2174/092986709789578213>
45. Lee Y, von Gunten U (2012) Water Res 46(19):6177–6195. <https://doi.org/10.1016/j.watres.2012.06.006>
46. Papa E, Villa F, Gramatica P (2005) J Chem Inf Model 45(5):1256–1266. <https://doi.org/10.1021/ci0502121>
47. Toropova AP, Toropov AA, Martyanov SE, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2012) Chemom Intell Lab Syst 110(1):177–181. <https://doi.org/10.1016/j.chemolab.2011.10.005>
48. Toropova AP, Toropov AA, Benfenati E, Gini G (2011) Chemom Intell Lab Syst 105(2):215–219. <https://doi.org/10.1016/j.chemolab.2010.12.007>
49. Toropov AA, Toropova AP, Benfenati E (2010) Mol Divers 14(1):183–192. <https://doi.org/10.1007/s11030-009-9156-6>
50. Toropov AA, Toropova AP, Kudyshkin VO (2022) Struct Chem 33(2):617–624. <https://doi.org/10.1007/s11224-021-01875-y>
51. Sivaraman N, Srinivasan TG, Vasudeva Rao PR, Natarajan R (2001) J Chem Inf Comput Sci 41(4):1067–1074. <https://doi.org/10.1021/ci010003a>
52. Toropov AA, Rasulev BF, Leszczynska D, Leszczynski J (2008) Chem Phys Lett 457(4–6):332–336. <https://doi.org/10.1016/j.cplett.2008.04.013>
53. Villaverde JJ, Sevilla-Morán B, López-Goti C, Alonso-Prados JL, Sandín-España P (2020) In: Shukla V, Kumar N (eds) Environmental concerns and sustainable development, air, water and energy resources, vol 1. Springer, Singapore, pp 1–27
54. Stoudmann N, Nowack B, Som C (2019) Environ Sci Nano 6(8):2520–2531. <https://doi.org/10.1039/c9en00472f>
55. Chopra SS, Bi Y, Brown FC, Theis TL, Hristovski KD, Westerhoff P (2019) Environ Sci Nano 6(11):3256–3267. <https://doi.org/10.1039/c9en00603f>
56. Organization for Economic Co-operation and Development (OECD) (2014) Ecotoxicology and environmental fate of manufactured nanomaterials. In: Series on the safety of manufactured nanomaterials, ENV/JM/MONO(2014)1, No. 40. OECD, Paris. Accessed 12 Aug 2022
57. Organization for Economic Co-operation and Development (OECD) (2020) Guidance document for the testing of dissolution and dispersion stability of nanomaterials and the use of the data for further environmental testing and assessment strategies. In: OECD guidelines for the testing of chemicals, ENV/JM/MONO(2020)9, No. 318. OECD, Paris. Accessed 12 Aug 2022
58. Camacho J, Smilde AK, Saccenti E, Westerhuis JA (2020) Chemom Intell Lab Syst 196:103907. <https://doi.org/10.1016/j.chemolab.2019.103907>
59. Siao MD, Shen WC, Chen RS, Chang ZW, Shih MC, Chiu YP, Cheng C-M (2018) Nat Commun 9(1):1442. <https://doi.org/10.1038/s41467-018-03824-6>

60. Toropov AA, Toropova AP, Benfenati E, Leszczynska D, Leszczynski J (2009) *J Math Chem* 46(4):1232–1251. <https://doi.org/10.1007/s10910-008-9514-0>
61. Toropova AP, Toropov AA, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2011) *Mol Divers* 15(1):249–256. <https://doi.org/10.1007/s11030-010-9245-6>
62. Toropova AP, Toropov AA (2019) *J Mol Struct* 1182:141–149. <https://doi.org/10.1016/j.molstruc.2019.01.040>
63. Mashino T, Shimotohno K, Ikegami N, Nishikawa D, Okuda K, Takahashi K, Nakamura S, Mochizuki M (2005) *Bioorg Med Chem Lett* 15(4):1107–1109. <https://doi.org/10.1016/j.bmcl.2004.12.030>
64. Toropova AP, Toropov AA, Benfenati E (2019) *Fuller Nanotub Carbon Nanostruct* 27(10):816–821. <https://doi.org/10.1080/1536383X.2019.1649659>
65. Marchesan S, Da Ros T, Spalluto G, Balzarini J, Prato M (2005) *Bioorg Med Chem Lett* 15(15):3615–3618. <https://doi.org/10.1016/j.bmcl.2005.05.069>
66. Salahnejad M, Zolfonoun E (2013) *J Nanopart Res* 15(11):2028. <https://doi.org/10.1007/s11051-013-2028-0>
67. Yilmaz H, Rasulev B, Leszczynski J (2015) *Nanomaterials* 5(2):778–791. <https://doi.org/10.3390/nano5020778>
68. Salahnejad M (2015) *Curr Top Med Chem* 15(18):1868–1886. <https://doi.org/10.2174/1568026615666150506145017>
69. Toropov AA, Toropova AP (2015) *Chemosphere* 124(1):40–46. <https://doi.org/10.1016/j.chemosphere.2014.10.067>
70. Toropov AA, Toropova AP (2015) *Chemosphere* 139:18–22. <https://doi.org/10.1016/j.chemosphere.2015.05.042>
71. Toropova AP, Toropov AA (2015) *Mini Rev Med Chem* 15(8):608–621. <https://doi.org/10.2174/1389557515666150219121652>
72. Toropova AP, Toropov AA, Benfenati E, Leszczynska D, Leszczynski J (2010) *J Math Chem* 48(4):959–987. <https://doi.org/10.1007/s10910-010-9719-x>
73. Toropov AA, Toropova AP (2014) *Chemosphere* 104:262–264. <https://doi.org/10.1016/j.chemosphere.2013.10.079>
74. Toropov AA, Toropova AP, Veselinović AM, Veselinović JB, Nesmerak K, Raska I Jr, Duchowicz PR, Castro EA, Kudyshkin VO, Leszczynska D, Leszczynski J (2015) *Comb Chem High Throughput Screen* 18(4):376–386. <https://doi.org/10.2174/1386207318666150305125044>
75. Toropov AA, Rallo R, Toropova AP (2015) *Curr Top Med Chem* 15(18):1837–1844. <https://doi.org/10.2174/1568026615666150506152000>
76. Fjodorova N, Novič M, Venko K, Drgan V, Rasulev B, Türker Saçan M, Sağ Erdem S, Tugcu G, Toropova AP, Toropov AA (2022) *Comput Struct Biotechnol J* 20:913–924. <https://doi.org/10.1016/j.csbj.2022.02.006>
77. Weininger D (1988) *J Chem Inf Comput Sci* 28(1):31–36. <https://doi.org/10.1021/ci00057a005>
78. Lotfi S, Ahmadi S, Zohrabi P (2020) *Struct Chem* 31(6):2257–2270. <https://doi.org/10.1007/s11224-020-01568-y>
79. Chopdar KS, Dash GC, Mohapatra PK, Nayak B, Raval MK (2020) *J Biomol Struct Dyn*. <https://doi.org/10.1080/07391102.2020.1867643>
80. Achary PGR, Toropova AP, Toropov AA (2019) *Int Food Res J* 122:40–46. <https://doi.org/10.1016/j.foodres.2019.03.067>
81. Pogány P, Arad N, Genway S, Pickett SD (2019) *J Chem Inf Model* 59(3):1136–1146. <https://doi.org/10.1021/acs.jcim.8b00626>
82. Fatemi MH, Malekzadeh H (2015) *J Iran Chem Soc* 12(3):405–412. <https://doi.org/10.1007/s13738-014-0497-4>
83. Toropova AP, Toropov AA, Rasulev BF, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2012) *Struct Chem* 23(6):1873–1878. <https://doi.org/10.1007/s11224-012-9996-z>
84. Toropov AA, Toropova AP, Martyanov SE, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2012) *Chemom Intell Lab Syst J* 112:65–70. <https://doi.org/10.1016/j.chemolab.2011.12.003>

85. Toropov AA, Toropova AP, Martyanov SE, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2011) *Chemom Intell Lab Syst* 109(1):94–100. <https://doi.org/10.1016/j.chemolab.2011.07.008>
86. Toropov AA, Benfenati E (2007) *Comput Biol Chem* 31(1):57–60. <https://doi.org/10.1016/j.combiolchem.2007.01.003>
87. Toropova AP, Toropov AA, Veselinović AM, Veselinović JB, Benfenati E, Leszczynska D, Leszczynski J (2016) *Ecotoxicol Environ Saf* 124:32–36. <https://doi.org/10.1016/j.ecoenv.2015.09.038>
88. Toropova AP, Toropov AA, Rallo R, Leszczynska D, Leszczynski J (2016) *Int J Environ Res* 10(1):59–64
89. Ahmadi S (2020) *Chemosphere* 242:125192. <https://doi.org/10.1016/j.chemosphere.2019.125192>
90. Cassano A, Robinson RLM, Palczewska A, Puzyn T, Gajewicz A, Tran L, Manganelli S, Cronin MTD (2016) *ATLA Altern Lab Anim* 44(6):533–556. <https://doi.org/10.1177/026119291604400603>
91. Toropov AA, Toropova AP, Leszczynska D, Leszczynski J (2019) *BioSystems* 181:51–57. <https://doi.org/10.1016/j.biosystems.2019.04.008>
92. Toropova AP, Toropov AA, Rallo R, Leszczynska D, Leszczynski J (2015) *Ecotoxicol Environ Saf* 112:39–45. <https://doi.org/10.1016/j.ecoenv.2014.10.003>
93. Toropova AP, Toropov AA, Manganelli S, Leone C, Baderna D, Benfenati E, Fanelli R (2016) *NanoImpact* 1:60–64. <https://doi.org/10.1016/j.impact.2016.04.003>
94. Achary PGR, Begum S, Toropova AP, Toropov AA (2016) *Mater Discov* 5:22–28. <https://doi.org/10.1016/j.md.2016.12.003>
95. Toropov AA, Achary PGR, Toropova AP (2016) *Chem Phys Lett* 660:107–110. <https://doi.org/10.1016/j.cplett.2016.08.018>
96. Toropov AA, Kjeldsen F, Toropova AP (2022) *Chemosphere* 303:135086. <https://doi.org/10.1016/j.chemosphere.2022.135086>
97. Ahmadi S, Aghabeygi S, Farahmandjou M, Azimi N (2021) *Struct Chem* 32(5):1893–1905. <https://doi.org/10.1007/s11224-021-01748-4>
98. Toropova AP, Toropov AA, Leszczynski J, Sizochenko N (2021) *Environ Toxicol Pharmacol* 86:103665. <https://doi.org/10.1016/j.etap.2021.103665>
99. Toropov AA, Toropova AP (2021) *Sci Total Environ* 772:145532. <https://doi.org/10.1016/j.scitotenv.2021.145532>
100. Toropova AP, Toropov AA, Leszczynska D, Leszczynski J (2021) *Comput Biol Med* 136:104720. <https://doi.org/10.1016/j.combiomed.2021.104720>

Chapter 15

Employing Quasi-SMILES Notation in Development of Nano-QSPR Models for Nanofluids



Kimia Jafari and Mohammad Hossein Fatemi

Abstract Nowadays, variant strategies are proposed and evaluated to find the best scenario for upgrading the high-accurate QSAR/QSPR modeling, particularly on nano-scale. One of the most interesting samples is nanofluids because of high potential in heat transfer applications. In the case of nano-QSPR, some optimum empirical conditions and characteristic features (e.g., size of nanoparticles and temperature) play impressive roles in nanofluids' properties. Quasi-simplified molecular input-line entry-system (quasi-SMILES) is nominated as valuable linear notation to meet the demands for representation of nanofluids, either chemical structure or defined conditions. The outcomes of nano-QSPR modeling of nanofluids by quasi-SMILES not only make possible the incorporation of molecular structure with experimental conditions in modeling process but also reveal the influence of some molecular features on studied thermophysical properties. Herein, recent studies on the development of predictive models of nanofluids using quasi-SMILES, which is a new trend to estimate the properties of nanofluids, were discussed comprehensively. It is remarkable to point out that the statistical evaluation of proposed models confirmed the predictability power, reliability, and credit of developed models in all reported cases. It is rational that scholars are working on improving QSAR/QSPR modeling; employing quasi-SMILES is an open opportunity to overcome the limitations of conventional molecular representation.

Keywords Quasi-SMILES · Nano-QSPR · Nanofluids · Thermophysical properties · CORAL software

K. Jafari · M. H. Fatemi (✉)
Chemometrics Laboratory, Faculty of Chemistry, University of Mazandaran, Babolsar, Iran
e-mail: mhfatemi@umz.ac.ir

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
A. P. Toropova and A. A. Toropov (eds.), *QSPR/QSAR Analysis Using SMILES and Quasi-SMILES*, Challenges and Advances in Computational Chemistry and Physics 33, https://doi.org/10.1007/978-3-031-28401-4_15

373

Nomenclature

Abbreviations

ANN	Artificial neural network
ANFIS	Adaptive neuro-fuzzy inference system
F_k	Extracted feature of quasi-SMILES
AAD	Average absolute deviation
CORAL	Correlation and logic
$CW(F_k)$	Correlation weight of F_k
CCC	Concordance correlation coefficient
C_p	Isobaric heat capacity
CII	Correlation intensity index
DTR	Decision tree regression
DCW	Optimal descriptor based on quasi-SMILES
EG	Ethylene glycol
IIC	Index of ideality of correlation
GBR	Gradient boosting regression
MAE	Mean absolute error
MLP	Multi-layer perceptron
Q^2	Leave-one-out cross-validated correlation coefficient
QSAR	Quantitative structure–activity relationship
QSPR	Quantitative structure–property relationship
Quasi-SMILES	Quasi-simplified molecular input-line entry-system
R^2	Correlation coefficient
RBF	Radial basis function
RFR	Random forest regression
RMSE	Root mean square error
LDM	Liquid drop model
LSSVM	Least square support vector machine
SVR	Support vector regression
TC	Thermal conductivity
TF	Target function

Greek Symbols

ρ	Density
φ	Volume fraction of nanoparticle (%)

Subscripts

bf	Base fluid
nf	Nanofluid
p	Nanoparticle
v	Volume fraction

Chemical Formula

Ag	Silver
Al ₂ O ₃	Aluminium oxide
AlN	Aluminum nitride
Au	Gold
Bi ₂ O ₃	Bismuth (III) oxide
CeO ₂	Cerium (IV) oxide
Co ₃ O ₄	Cobalt (II,III) oxide
Cr ₂ O ₃	Chromium (III) oxide
Cu	Copper
CuO	Copper oxide
Dy ₂ O ₃	Dysprosium (III) oxide
Fe	Iron
Fe ₂ O ₃	Iron (III) oxide
Fe ₃ O ₄	Iron (II,III) oxide
Gd ₂ O ₃	Gadolinium (III) oxide
HfO ₂	Hafnium (IV) oxide
In ₂ O ₃	Indium (III) oxide
La ₂ O ₃	Lanthanum oxide
MgO	Magnesium oxide
Mn ₂ O ₃	Manganese (III) oxide
Mn ₃ O ₄	Manganese (II,III) oxide
Ni ₂ O ₃	Nickel (III) oxide
NiO	Nickel (II) oxide
Sb ₂ O ₃	Antimony oxide
Si ₃ N ₄	Silicon nitride
SiC	Silicon carbide
SiO ₂	Silicon dioxide
SnO ₂	Tin (IV) oxide
TiN	Titanium nitride
TiO ₂	Titanium dioxide
WO ₃	Tungsten (VI) oxide
Y ₂ O ₃	Yttrium (III) oxide
Yb ₂ O ₃	Ytterbium (III) oxide

ZnO	Zinc oxide
ZrO ₂	Zirconium oxide

15.1 Introduction

15.1.1 Nanofluids

Nowadays, the sustainability plan on a global scale is toward the upgrade of energy efficiency in various scopes, for instance, high-efficient cooling systems, which are common in automobile radiators and air conditioning [1, 2]. One of the most interesting novel products of nanotechnology is nanofluids, which are of broad potential for implementations in heat transfer operations and thermofluid systems by a remarkable enhancement in their performance. The suspension of nano-scale materials (including nanoparticle, nanotube, and nano-rod) with 1–100 nm size range in a conventional base fluid such as water, ethylene glycol, transfer oil, ionic liquids, and deep eutectic solvents are named nanofluids [3, 4], which cause to notable progress in thermophysical properties. Taking into account the unique characteristics of nanofluids, in particular viscosity and thermal conductivity, it is predictable to take the place of traditional options in heat transfer equipment. Accordingly, these materials are a hot topic in both scientific researches and industrial applications.

Nanofluids are prepared in two main ways: single-step and two-step methods. Most privileges belong to the two-step method since it is user-friendly and ease-doing. Nanostructures (nanoparticles, carbon nanotubes, graphene, etc.) are added to a liquid fluid and then mixed properly by a mechanical homogenizer up to form a uniform suspension. With respect to this issue that homogeneity of nanofluids should be considered as a crucial aspect since it will influence thermophysical effectiveness, different processes are suggested to stabilize prepared nanofluids such as microwave radiation, usage of electrochemical equipment, and ultra-sonication (which is the most common method) [4, 5]. In the last years, there have been paramount contributions, mainly focused on the design of new nanofluids (considering various nanoparticles in mono or hybrid forms and different base fluids), well-dispersion strategies, and ideas to use in numerous scopes (such as cooling, lubricant, and phase change materials) [2, 6–8]. Moreover, regarding the high potential of employing nanofluids for heat transfer operations, very diverse papers are published in respect to the evaluation of thermophysical properties especially thermal conductivity and viscosity by different types of nanofluids, variant nanoparticles' concentration, and temperature. Some effective parameters on thermal conductivity, as the most noteworthy property, are shown in Fig. 15.1.

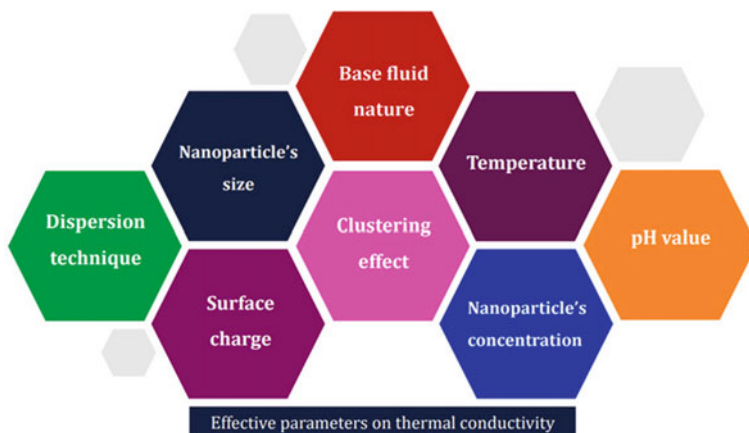


Fig. 15.1 Some of the most impressive parameters on thermal conductivity of nanofluids

15.1.2 Theoretical Methods Applied for Study of Nanofluids' Properties

Despite there being many and ever-increasing experimental studies focused on the introduction of new nanofluids, measurement of their thermophysical properties, and analysis of their applications, the theoretical studies on a survey of the effective features of nanofluids' characteristics in relevant literature are still limited (Fig. 15.1). Study of nanofluids with a theoretical point of view is pivotal, not only to subtract the quantity of high-cost experimental inspections (which are energy/time-consuming as well), but to appreciate also the impressive components on nanofluids' thermophysical properties.

Among all available theoretical approaches, artificial neural network (ANN), radial basis function (RBF), adaptive neuro-fuzzy inference system (ANFIS), and support vector machine (SVM) are the most usual techniques, which are utilized to develop models for properties of nanofluids [8, 9]. For example, Sharma et al. [10] were collected a data set including 228 experimental thermal conductivity values of TiO_2 dispersed in water with different sizes and shapes, then modeled thermal conductivity of nanofluids by five algorithms, ANN, support vector regression (SVR), random forest regression (RFR), gradient boosting regression (GBR), and decision tree regression (DTR) algorithms. Eventually, gradient boosting was suggested as the best algorithm with precise analysis and confirmed that the shape of titania nanoparticles affected the thermal conductivity predictions of the nanofluids. In another study, Cui et al. [11] studied the effective parameters on thermal conductivity of nanofluids experimentally and theoretically. An empirical data set contains 469 data (80 collected from their experiments, 389 collected from relevant literature) which were considered of TiO_2 nanoparticles (in shapes sheet, spherical, clubbed, and ellipsoidal) suspended in water in temperature range of 20–60 °C. In order to generate a

predictive network, four factors including shape factor, thermal conductivity, and concentration of nanoparticles and temperature introduced as input and thermal conductivity of nanofluids defined as a single output. Then, RBF, LSSVM, ANFIS, multi-layer perception (MLP), generalized regression neural network, and cascade feedforward were used to estimate statistical performance. Ultimately, the cascade feedforward neural network trained by Levenberg–Marquardt was nominated as the best-optimized network.

15.1.3 The Importance of QSPR Study for Nanofluids

Taking into account, great attention to nanofluids has been received both from the scientific community and industry, and the fact that their fabrication may increase by the year, environmental effects of these materials should be addressed, since such effects can accumulate and spread. Furthermore, revision of relevant literature indicated despite that the molecular structure of nanoparticles is impressment on nanofluids' properties, particles' size, volume fraction of nanoparticles and temperature are the most popular variables in theoretical studies [12]. Hence, the development of credible procedures to forecast their properties and/or activities connecting with the molecular structure is a serious object.

A well-known powerful paradigm to design an accurate mathematical model related to the physicochemical properties and structural features of compounds is a quantitative structure–property relationship (QSPR), which is entitled nano-QSPR when performed for nano-scale samples [13–16]. Although there are various challenges in the route of chemometric studies of nanofluids which are arising from that the structure elucidation of nanofluids is still an open question, some nano-QSPR models generation has been successfully accomplished. Herein, besides the relevant insights in the current subject and providing prominent case studies, some key strategies used in scientific reports to reach successful usage of quasi-simplified molecular input-line entry-system (quasi-SMILES) notation in the development of nano-QSPR models are discussed in detail. The future directions on using quasi-SMILES in designing predictive models for nanofluids are given as well.

15.2 Methodology of CORAL-Based Models Generation

15.2.1 Collection of a Valid Data Set

The quality of QSPR procedure strongly appertains to the gathered experimental data set. Therefore, collecting a consistent and reliable data set is decisive. Since erratic data lift the risk of inconstant modeling process and then generate QSPR

models with deficient statistical performance and/or predictability, a standard guideline including key criteria should apply to appraise existing data. Depending on the final purpose of generating nano-QSPR models, the intended principles toward collecting an acceptable data set are as follows:

- (i) accessible in a proper volume;
- (ii) supported with adequate concerned information;
- (iii) performed by good-quality laboratory skill through a clarified protocol; and
- (iv) confirmed by sufficient chemical characterization tests [17].

Also, this fact should be highlighted that different testing conventions, which are enforced by discrete operators/laboratories may cause a notable variance in data. Thus, although broad diversity is more suitable in collecting data, it is momentous to keep a logical coherency in selected data with regard to their method of analysis as much as possible.

Fortunately, the rapid growth of empirical examinations of nanofluids' thermo-physical properties is highly profitable for complementary theoretical projects. The number and quantity of data for different properties, in particular thermal conductivity and viscosity, are not restricted and properly available in scientific journals. In order to gather a precise data set to start nano-QSPR modeling of nanofluids, the type, size, shape, and concentration of nanoparticles, base fluid nature, preparation method, temperature, and any specific experimental conditions in regard to the studied target property should define accurately.

15.2.2 Quasi-SMILES for Nanofluids

The quasi-SMILES is the advanced version of SMILES string, which is nowadays recognizing as a promising tackle to import molecular representation, especially in QSPR and nano-QSPR studies. Overall, quasi-SMILES is constructed of two key components, the first part includes chemical composition of the studied sample clarified by SMILES structure, and the other part consists of a symbols chain, which are codes of all available/intended experimental conditions and/or possible complementary information to better describe target samples [16, 18, 19].

The molecular structure of nanofluids is in such a way that besides chemical representation of nanoparticles, it is necessary to consider extra variables to represent them accurately. The diversity of nanofluids is actually wide, changing nanoparticles and base fluid form a new nanofluids, and it can be even much more when two elements are defined for each part, such as hybrid nanofluids. Along with variety in possible choices as nanoparticles (different types, different size, shape, and concentration) and base liquid (such as water, organic solvents, ionic liquids, and deep eutectic solvents), which create a novel nanofluid, the specified experimental conditions (e.g., preparation method, temperature, pressure, dispersion technique, sonication time) are leading to define a new sample in nano-QSPR studies. In this condition,

quasi-SMILES notation is the best option to reflect nanofluids. Indeed, using quasi-SMILES makes possible to generate predictive models for any planned thermophysical properties of nanofluids while equation is depended on each feature/condition that encoded in quasi-SMILES structure. It should be mentioned that more defined details in quasi-SMILES notation would cause a more informative model as an outcome, which is truly valuable. In order to display a proper outlook of the concept of quasi-SMILES structure for nanofluid samples, Fig. 15.2 is provided.

15.2.3 Optimal Descriptors, Predictability Criteria, and Optimization

The whole modeling process is developed using the CORAL free and open-access framework. The optimal descriptors based on extracted features of intended quasi-SMILES calculated as the following:

$$DCW(T^*, N^*) = \sum_{k=1}^n CW(F_k) \quad (15.1)$$

where T^* and N^* are the specific parameters of Monte Carlo method, while $CW(F_k)$ is the calculated correlation weight for each certain quasi-SMILES feature [13, 16, 20, 21]. Using the calculated optimal descriptors, a one variable correlation would be developed as overall format of suggested model, as follows:

$$\text{Target property} = C_0 + C_1 \times DCW(T^*, N^*) \quad (15.2)$$

where C_0 and C_1 are regression coefficients. The index of ideality of correlation (IIC) and correlation intensity index (CII) is suggesting as possible tools to rectify the predictive power of developed models. The detail of their calculation along with complete modeling process was well discussed in literature [13, 14, 16, 20]. It is notable to point out that precise statistical appraisalment is a drastic step in model development. Through Monte Carlo method, some common criteria such as R^2 , CCC, Q^2 , and R_m^2 parameters are employed to estimate the credit and validity of generated models.

15.3 Successful Nano-QSPR Studies on Nanofluids

This section is specialized in reviewing recent trend researches in nano-QSPR modeling of nanofluids. With the aim of covering all available studies on nano-QSPR modeling of nanofluids, besides performed papers with a precise focus on the elucidation of the underlying physical aspects of these substances by quasi-SMILES, a brief

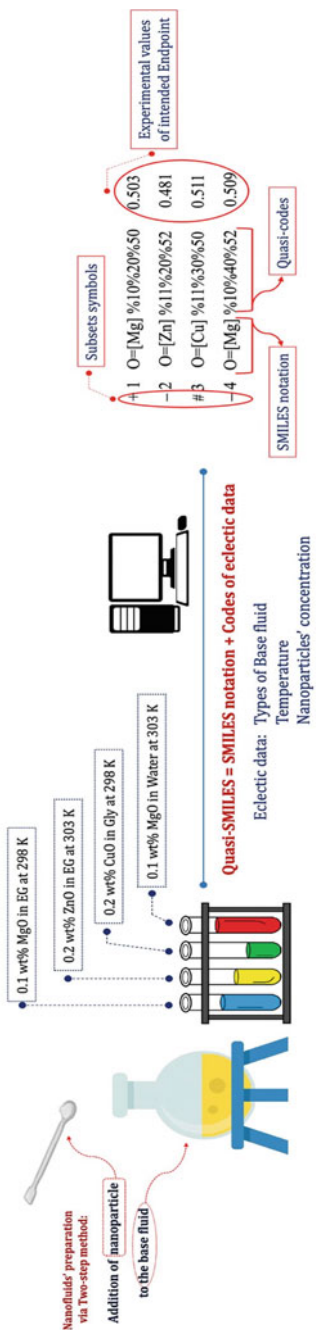


Fig. 15.2 Graphical representation of nanofluids and their related quasi-SMILES structure

overview of developed models for nanofluids using different methods is presented as well.

To the best of one's knowledge, the first paper on QSPR study of nanofluids has been performed by Sizochenko et al. [22] in 2015. They defined the assessment of the liquid drop approach to the model thermal conductivity of nanofluids as the main object of the project. Taking into account that thermal conductivity is dependent on both size and shape of nanoparticles, the domain of application was confined to spherical and near-spherical nanoparticles, and in addition to volume fraction and size of nanoparticles, liquid drop model (LDM)-based descriptors were applied as size-dependent descriptors for a series of nanoparticles. Toward investigating the effectual structural features of nanofluids on the improvement of thermal conductivity, random forest regression was applied, which is a stepwise non-parametric method. In the end, a model composed of 10 trees was developed.

After the aforementioned paper, a detailed analysis of viscosity and thermal conductivity of nanofluids was performed by Sizochenko et al. [23]. A varied database includes silica-derived, metals, and metal oxides nanoparticles with diverse size, and concentrations dispersed in water at the common range of temperature (21–25 °C) were subjected to modeling. Weka software package was used for computational process, and M5P classifier, which makes possible to incorporate a decision tree model with linear regressions at nodes, was exerted. It should be stated that structure of nanofluids was reflected by a new suggested hierarchical combination system of descriptors such as thickness and concentration of interfacial layer and size of nanoparticles. The final proposed model for viscosity was assigned by $R^2 = 0.79$ and $RMSE = 0.234$; also for thermal conductivity, it was determined by $R^2 = 0.81$ and $RMSE = 0.055$. Moreover, the increment in interfacial layer thickness, surface area ratio, and weighted fraction-dependent factors, together with decrement in size of nanoparticles were reported as efficient parameters on raising thermal conductivity and viscosity of nanofluids.

Sizochenko et al. [24] applied a simple case of linear regression to build theoretical models for thermal conductivity of nanofluids based on alumina and copper oxide. In order to represent dependency of target property to size and concentration of nanoparticles, the weighted surface-area-to-volume ratio descriptor was calculated. Regarding their findings were confirmed that the dependency of thermal conductivity on concentration and size of nanoparticles was non-linear for alumina-based nanofluids and linear for CuO-based ones, and rising in weighted fraction-dependent parameters lead to thermal conductivity enhancement.

With the aim of focusing on building up predictive models for nanofluids using quasi-SMILES, a detailed and comprehensive review of recently published studies is provided as follows. The stability of nanofluids is such an important aspect, which should be followed precisely due to its direct influence on thermophysical properties. Evaluation of zeta potential is one of the most usual analyzes to remark the stability of nanofluids, which Toropov et al. [25] chose it as an endpoint to subject nano-QSPR modeling using quasi-SMILES.

Eighty-seven zeta values of metal oxide nanoparticles in water were compiled, then nominal size and size in the medium were ascertained as basic attributes which

all encoded by specialized quasi-SMILES. Some example of defined quasi-SMILES structures in various studies is represented in Table 15.1.

Considering three different splits, mathematical models with content statistical qualities were built up, while they mentioned that because of high deviation, the obtained models were restricted to the aim of stability prediction. Nevertheless, through three different target functions, Toropov et al. [20] performed extra calculations on the previous benchmark data set to provide a comprehensive model evolution in zeta potential of nanofluids. The preferable target functions were defined with respect to different predictive ability criteria as follows:

$$\text{Target function \#1: } TF_1 = R + R' - |R - R'| \times 0.1 \quad (15.3)$$

$$\text{Target function \#2: } TF_2 = TF_1 + IIC \times 0.2 \quad (15.4)$$

$$\text{Target function \#3: } TF_3 = TF_1 + CII \times 0.2 \quad (15.5)$$

where R is correlation coefficient for training set and R' is that one for invisible training set. Each target function was checked by three different runs, and reported statistical qualities were quite satisfying since at the best result $R_{\text{validation}}^2 = 0.9336$ and $RMSE_{\text{validation}} = 6.6$ while the best developed model in the previous study reached $R_{\text{validation}}^2 = 0.8213$ and $RMSE_{\text{validation}} = 15.8$.

Due to the high potential application of nanofluids in heat transfer systems, thermal conductivity is the exact property that attracts the most attention. Some reputable theoretical relationships of thermal conductivity used repetitively in relevant literature are as follows, which revealed thermal conductivity of all elements (nanoparticles, base fluids, and nanofluids), and volume fraction is the common variables.

Maxwell [27]

$$k_{\text{eff}} = k_f \frac{k_p + 2k_f + 2\varphi(k_p - k_f)}{k_p + 2k_f - \varphi(k_p - k_f)} \quad (15.6)$$

Bruggeman [27]

$$\varphi \left[\frac{k_p - k_{\text{eff}}}{k_p + 2k_{\text{eff}}} \right] + (1 - \varphi) \left[\frac{k_f - k_{\text{eff}}}{k_f + 2k_{\text{eff}}} \right] = 0 \quad (15.7)$$

Hamilton and Crosser [27]

$$k_{\text{eff}} = k_f \frac{k_p + (n - 1)k_f + (n - 1)\varphi(k_p - k_f)}{k_p + (n - 1)k_f - \varphi(k_p - k_f)} \quad (15.8)$$

Table 15.1 Used quasi-SMILES molecular representation in different studies

Study	Eclectic data	Studied range	Quasi-code	Example of quasi-SMILES	Meaning
Toward the development of global nano-QSPR models: zeta potentials of metal oxide nanoparticle [25]	Nominal size (nm)	3.59–420.0	%11–%40	O=[Al]O[Al]=O%11%51	Al ₂ O ₃ nanoparticle with a nominal size of 11.4 nm and media size of 94.7 nm
	Size in H ₂ O (nm)	28.90–6000.0	%51–%80		
Application of nano-QSPR paradigm to develop predictive models for thermal conductivity of metal oxide-based ethylene glycol nanofluids [13]	Volume fraction (%)	0.2–7.00	%31–%53	O=[Zn]%22%33%61	ZnO with double bond and particle size of 48 nm in 0.79–1.08 v.% at 10 °C
	Temperature (°C)	10–70	%61–%73		
	Size (nm)	5–60	%11–%24		
A new approach to model isobaric heat capacity and density of some nitride-based nanofluids using Monte Carlo method [14]	Temperature (K)	288.15–305.15	%20–%22	N#[Al]%10%20%30	AlN with triple bond, size of 20 nm in 0.01 wt.% at 288.15 K
	Size (nm)	20–80	%10–%13		
	Mass fraction (%)	0.01–0.1	%30–%33		
The development of nano-QSPR models for viscosity of nanofluids using the index of ideality of correlation and the correlation intensity index [26]	Size (nm)	20–100	%10–%14	1. O=[Cu]%11%31 2. O=[Si]=O%20%31	1. CuO with size of 40 nm and 2 v.% concentration 2. SiO ₂ with blade shape and 2 v.% concentration
	Volume fraction (%)	1–5	%30–%34		
	Shape	5 different shapes	%20–%24		

At the first try of implementation of quasi-SMILES as a tool to generate models for thermal conductivity of nanofluids, Jafari and Fatemi [13] collected a reliable data set involving several common-use metal oxides nanoparticles in ethylene glycol, the second choice as the popular base liquid. As far as is known, this is the most general and largest provided data set in nano-QSPR studies of nanofluids up to now. Four random split were designated to build up models, which averagely a training sets of 270 nanofluids and a validation set of 90 nanofluids were used. Monte Carlo optimization has provided an interesting option, named promoters, which are the sole features extracted of quasi-SMILES with reiterative positive/negative computed CW values in all executed runs. Regarding the sign of CW for each promoter, an increasing or decreasing effect on the endpoint would be granted to those features. The authors reported by checking calculated CWs in all splits, it was concluded that high volume fractions and nanoparticles size of 20 and 31 nm had a positive role on thermal conductivity, while the feature represented of double bond and low volume fractions (in the range of 0.2–0.75) displayed a reduction impact on thermal conductivity. Table 15.2 represents a nutshell of some considerable impressive features extracted by Monte Carlo modeling using quasi-SMILES.

In order to make proposed models for nanofluids thermal conductivity more acceptable, Jafari et al. [28] regenerated nano-QSPR model by a specific index, CII, and recommended new models using same original data set of ethylene glycol-based metal oxide nanofluids and identical data distribution to sub-sets defined in the previous study [13]. Their theoretical findings have confirmed that applying CII in target function cause a considerable augment in statistical characteristics of developed models and built up more robust computational relationships since the range of correlation coefficient and leave-one-out cross-validated coefficient (Q^2) of validation set, respectively, in the previous study [13] were achieved 0.68–0.86 and 0.66–0.85, while by concerning CII they improved up to $R^2_{\text{validation set}} = 0.82\text{--}0.91$, and $Q^2_{\text{validation set}} = 0.81\text{--}0.90$. Also, in the case of mean absolute error (MAE) of

Table 15.2 Some highlighted effective structural features on thermophysical properties of nanofluids extracted by developed nano-QSPR models

Properties of nanofluids	Notable features	
	Positive	Negative
Thermal conductivity	High φ_v , T range (24–55 °C), Al, Ce, Mg, Ti, Zn [13]	Low φ_v (0.2–0.79 v.%), double bond [13]
Zeta potential	Low nominal size range (3.59–45.23 nm), and Ni [25]	Not defined [25]
Density	Size range (25–50 nm), Al, N, Ti [14]	Double bond, triple bond, low φ_v (0.01 wt.%) [14]
Isobaric heat capacity	Low φ_v (0.01–0.05 wt.%), size range (20–80 nm), Al, Ti [14]	Double bond, high φ_v (0.1 wt.%) [14]
Viscosity	High φ_v (3–5 v.%), O, Zn, shapes of platelet, cylindrical [26]	Low φ_v (1 v.%), Al [26]

validation set in the previous study [13] was calculated in the range of 0.028–0.037, while via using suggested criteria, CII, it decreased to the range of $MAE_{\text{validation set}} = 0.023\text{--}0.029$.

In the route of evaluation of nanofluids' characteristics, Jafari and Fatemi [14] investigated the modeling of density and isobaric heat capacity of some nitride-based nanofluids using the Monte Carlo method in CORAL framework. Even though the thermophysical properties of dispersions containing nano-sized particles have been extensively studied in the literature, there is still a lack of accurate models to predict or correlate these kinds of properties. The optimal descriptors based on quasi-SMILES considering chemical structure of AlN, TiN, and Si₃N₄, as well as temperature, size, and concentration of nanoparticles were computed, and predictive models were generated via Monte Carlo optimization for three random splits in each property. To take a closer look at proficiency of suggested models by way of nano-QSPR paradigm, a comparison with popular classic models was subjected to gage. Pak and Cho [29] is a classic model to describe the heat capacity of nanofluids, which the equation is as follows:

$$C_{p,nf} = (1 - \varphi_v)C_{p,bf} + \varphi_v C_{p,p} \quad (15.9)$$

Moreover, in order to illustrate the relation of the density of nanofluids, Pak and Cho [29] equation is given below:

$$\rho_{nf} = (1 - \varphi_v)\rho_{bf} + \varphi_v \rho_p \quad (15.10)$$

Jafari and Fatemi [14] compared qualities of acquired models by Monte Carlo method to the aforementioned classical equations. Interestingly, the efficiency of proposed models was premier to the classical models since statistical characteristics of the best split (split 1) were calculated as $R^2 = 96.8$ and $AAD = 0.225$ for density whenever by Pak and Cho equation was as $R^2 = 96.3$ and $AAD = 0.302$, also in the case of isobaric heat capacity as the best results of nano-QSPR models (split 2), $R^2 = 96.8$ and $AAD = 0.447$, and the worst results were calculated by split 1 with $R^2 = 93.2$ and $AAD = 0.640$, while using Pak and Cho equation $R^2 = 86.0$ and $AAD = 3.593$. It was crystal clear that even by worst split, the performance of developed nano-QSPR models was better than classic equations. Furthermore, the outcomes obtained by calculated CWs made disclosed that some attributes such as double and triple bond influence on density and isobaric heat capacity of nanofluids, while nanoparticles' size did not efficient touch on intended properties.

Lately, Jafari et al. [26] used quasi-SMILES representation to take into consideration the size and shape of nanoparticles in modeling nanofluids' viscosity by calculation of optimal descriptors through Monte Carlo method. Their contribution not only had been directed to the calculation of size and shape-dependent optimal descriptors but also supplied a comparison of model generation using different indexes, CII and IIC, simultaneously. An authentic data sets contain four types of nanoparticles suspended in water which was distinguished with the aim of survey the size effect. Also, another data set of 100 water-based nanofluids was utilized to take into account

the particles' shape effect on viscosity. In order to division of the original data sets, the authors maintained the ratio of 25% for all training, invisible training, calibration, and validation sets. The authors asserted that although the results achieved from three random splits in both intended target functions (using CII and IIC) were credible and robust for total sub-sets, the statistical qualities have markedly elevated when the CII was included in the target function. For instance, the reported results for the best achieved splits were as follows: for data set I (study of size effect), R^2 value of split 1 was increased from 0.8686 to 0.9444, and for data set II (study of shape effect), R^2 value of split 1 was enhanced from 0.8230 to 0.9402.

This chapter dealt with an overview of nanofluids, their characteristics, and theoretical studies in particular by QSPR paradigm (Table 15.3).

15.4 Conclusion and Perspective Outlook

A general modeling workflow based on optimal descriptor of quasi-SMILES was subjected to provide an overall outlook of the development nano-QSPR modeling of nanofluids, which did not require long-term and complicated computations. Owing to the simplicity, transparency, and availability of empirical data, it is expected even if just a few studies have been currently reported in applying quasi-SMILES in the design of nano-QSPR models for thermophysical properties of nanofluids, the attention on this notation due to high potential in considering of different aspects of nanofluids would be continuously growing. It was discussed that the proposed models on nanofluids' properties not only should be statistically sound, trusty, and robust, but it is better also consist of varied data sets in order to have a satisfying applicability domain. By a comprehensive review of relevant literature, a number of nano-QSPR developed models based on quasi-SMILES for the most prominent thermophysical properties of nanofluids, i.e., thermal conductivity, density, isobaric heat capacity, and viscosity were discussed. It was mentioned accurately, the proposed models on nanofluids' properties not only should be statistically sound, trusty, and robust, but it is better also consist of varied data sets in order to have a satisfying applicability domain. Furthermore, regarding the successful studies, it was confirmed the models generated by the application of CII are statistically more valid than those developed with IIC. It worth to mention that the newest version of CORAL software is offered novel target functions based on CII, which could be profitable to develop predictive nano-QSPR models for various thermophysical properties of nanofluids with possible better statistical performance.

Since moving forward always should be appreciated, the availability of all data sets utilized in reviewed studies (well provided by authors in manuscripts and/or supplementary materials) encourages other chemometric scholars to challenge the current suggested method by performing further researches on a generation of theoretical models for nanofluids and supply a competition between nano-QSPR models based on quasi-SMILES with other possible ones. In spite of the fact that it is explicit that experimental strategies can never be quite replaced by computational process,

Table 15.3 Summarized quantitative structure–property relationship (QSPR) studies on nanofluids

#	Author	Target property	Data set	Nanofluids	The best model	R^2	s	MAE	Comments
1	Sizochenko et al. [22]	Thermal conductivity	23	Fe, Al ₂ O ₃ , Fe ₂ O ₃ , Cu, TiO ₂ , ZrO ₂ in water	Model consisted of 10 trees with 5 descriptors	0.77	0.06	–	Liquid drop model-based descriptors used for nanoparticles with various sizes
2	Sizochenko et al. [23]	Thermal conductivity	100	Au, Al ₂ O ₃ , Ag, Fe ₂ O ₃ , Cu, ZrO ₂ , CuO, TiO ₂ , SiO ₂ , SiC, Fe in water	Models developed based on if–then–else rules, consisted of nine descriptors	0.81	0.055	–	Descriptors: heat capacity per ϕ_v , number of Nps per ϕ_v , number of 5th-period cations in chemical formula of Nps, presence/absence of metal oxides of periods 4 and 5, M_w , and density
3	Sizochenko et al. [23]	Viscosity	69	Au, Al ₂ O ₃ , Ag, Fe ₂ O ₃ , Cu, ZrO ₂ , CuO, TiO ₂ , SiO ₂ , SiC, Fe in water	Models developed based on if–then–else rules, consisted of seven descriptors	0.78	0.244	–	Descriptors: number of Nps per ϕ_v , surface ratio, number of 3rd-period cations in chemical formula of Nps, heat capacity per ϕ_v , interfacial nanolayer thickness, density per ϕ_v , and presence/absence of TiO ₂

(continued)

Table 15.3 (continued)

#	Author	Target property	Data set	Nanofluids	The best model	R^2	s	MAE	Comments
4	Sizochenko et al. [24]	Thermal conductivity	29	Al ₂ O ₃ in water	Suggested a decision tree with 3 rules	0.826	2.5	–	Data set of was defined in NanoBRIDGES framework
5	Sizochenko et al. [24]	Thermal conductivity	12	CuO in water	$k = 1773.58 F_w^{**} + 1.6122$	0.912	5.2	–	The model was developed based on three descriptors: volume fraction and size of nanoparticles, and weighted number of imaginary parts per volume
6	Jafari and Fatemi [13]	Thermal conductivity	360	ZnO, Al ₂ O ₃ , MgO, CeO ₂ , TiO ₂ , Fe ₂ O ₃ , Fe ₃ O ₄ , Co ₃ O ₄ , CuO, and SnO ₂ in EG	TCR = 0.8948 (\pm 0.001) + 0.0454 (\pm 0.0002) \times DCW(3, 21)	0.860	–	0.028	The largest and most diverse data set in thermal conductivity of nanofluids for nano-QSPR study was collected
7	Jafari et al. [28]	Thermal conductivity	360	ZnO, Al ₂ O ₃ , MgO, CeO ₂ , TiO ₂ , Fe ₂ O ₃ , Fe ₃ O ₄ , Co ₃ O ₄ , CuO, and SnO ₂ in EG	TCR = 1.2347 + 0.05586 \times DCW(1, 15)	0.910	0.030	0.023	The correlation intensity index (CII) employed as a new element in build-up models to improve power of prediction

(continued)

Table 15.3 (continued)

#	Author	Target property	Data set	Nanofluids	The best model	R^2	s	MAE	Comments
8	Toropov et al. [25]	Zeta potential	87	Al ₂ O ₃ , ZnO, Co ₃ O ₄ , MgO, Dy ₂ O ₃ , CeO ₂ , Cr ₂ O ₃ , Fe ₂ O ₃ , CuO, Bi ₂ O ₃ , Fe ₃ O ₄ , ZrO ₂ , Gd ₂ O ₃ , HfO ₂ , SiO ₂ , In ₂ O ₃ , La ₂ O ₃ , WO ₃ , Mn ₂ O ₃ , Mn ₃ O ₄ , TiO ₂ , Sb ₂ O ₃ , Ni ₂ O ₃ , NiO, Yb ₂ O ₃ , Y ₂ O ₃ in water	$\xi = 1.044 (\pm 0.524) + 13.666 (\pm 0.238) \times DCW(1, 30)$	0.821	15.8	11.6	Three models were developed based on specific quasi-SMILES descriptors, which were reflected size-dependent behavior of zeta potentials, using Monte Carlo method
9	Jafari and Fatemi [14]	Isobaric heat capacity	72	AlN, Si ₃ N ₄ , TiN in EG	$C_{p,r} = 0.8669 (\pm 0.0017) + 0.0105 (\pm 0.0002) \times DCW(1, 6)$	0.974	0.007	0.005	Three random splits were used to generate models, which were quite robust, without any outliers
10	Jafari and Fatemi [14]	Density	54	AlN, Si ₃ N ₄ , TiN in EG	$\rho_l = 0.9914 (\pm 0.0003) + 0.0046 (\pm 0.00006) \times DCW(3, 6)$	0.976	0.003	0.002	Three random models were developed, which all had R^2 and Q^2 higher than 0.8

(continued)

Table 15.3 (continued)

#	Author	Target property	Data set	Nanofluids	The best model	R^2	s	MAE	Comments
11	Toropov et al. [20]	Zeta potential	87	Al ₂ O ₃ , ZnO, Co ₃ O ₄ , MgO, Dy ₂ O ₃ , CeO ₂ , Cr ₂ O ₃ , Fe ₂ O ₃ , CuO, Bi ₂ O ₃ , Fe ₃ O ₄ , ZrO ₂ , Gd ₂ O ₃ , HfO ₂ , SiO ₂ , In ₂ O ₃ , La ₂ O ₃ , WO ₃ , Mn ₂ O ₃ , Mn ₃ O ₄ , TiO ₂ , Sb ₂ O ₃ , Ni ₂ O ₃ , NiO, Yb ₂ O ₃ , Y ₂ O ₃ in water	$\zeta = 31.92 (\pm 1.02) + 9.82 (\pm 0.18) \times DCW(1, 15)$	0.934	6.6	5.2	Usage of CII cause more robust and reliable models statistically
12	Jafari et al. [26]	Viscosity	100	Al ₂ O ₃ , CuO, ZnO, SiO ₂ in water	$\eta_r = 0.8887 (\pm 0.0089) + 0.2301 (\pm 0.0046) \times DCW(1, 15)$	0.944	0.074	0.061	Three random splits by 2 different target functions were modeled
13	Jafari et al. [26]	Viscosity	100	Al ₂ O ₃ , CuO, ZnO, SiO ₂ in water	$\eta_r = -0.9005 (\pm 0.0740) + 1.3813 (\pm 0.0314) \times DCW(1, 15)$	0.940	0.260	0.212	Study of particles' shape effect on viscosity was the main object

F_w is weighted surface-area-to-volume ratio, which is a LDM-based weighted descriptor

these approaches can be integrated to provide a better comprehension. Nevertheless, it is well expected to consider nano-QSPR based on quasi-SMILES as an exciting trend in theoretical studies of nanofluids due to great potential in introducing computational relationships of nanofluids' thermophysical properties with different aspects including structural features and experimental circumstances. Hence, one can conclude that the demand to develop more number of nano-QSAR models is not only advisable but also supports its certain role in prediction of nanofluids' characteristics.

Declaration of Competing Interest The authors declare that they have not any known personal relationships or competing financial interests that could have appeared to effect on this chapter.

References

1. Hamze S, Cabaleiro D, Estellé P (2021) *J Mol Liq* 325:115207. <https://doi.org/10.1016/j.molliq.2020.115207>
2. Ghalandari M, Maleki A, Haghighi A, Shadloo MS, Nazari MA, Tlili I (2020) *J Mol Liq* 313:113476. <https://doi.org/10.1016/j.molliq.2020.113476>
3. Jafari K, Fatemi MH, Estellé P (2021) *J Mol Liq* 321:114752. <https://doi.org/10.1016/j.molliq.2020.114752>
4. Asadi A, Aberoumand S, Moradikazerouni A, Farzad P, Gawel Ź, Patrice E, Omid M, Somchai W, Nguyen Hoang M, Arabkoohsar A (2019) *Powder Technol* 352:209–226. <https://doi.org/10.1016/j.powtec.2019.04.054>
5. Asadi A, Pourfattah F, Miklós Szilágyi I, Afrand M (2019) *Ultrason Sonochem* 58:104701. <https://doi.org/10.1016/j.ultsonch.2019.104701>
6. Hajatzadeh Pordanjani A, Aghakhani S, Afrand M, Mahmoudi B, Mahian O, Wongwises S (2019) *Energy Convers Manag* 198:111886. <https://doi.org/10.1016/j.enconman.2019.111886>
7. Rashidi S, Mahian O, Languri EM (2018) *J Therm Anal Calorim* 131:2027–2039. <https://doi.org/10.1007/s10973-017-6773-7>
8. Mahian O, Kolsi L, Amani M, Estellé P, Ahmadi G, Kleinstreuer C, Marshall JS, Siavashi M, Taylor RA, Niazmand H, Wongwises S (2019) *Phys Rep* 790:1–48. <https://doi.org/10.1016/j.physrep.2018.11.004>
9. Maleki A, Haghighi A, Mahariq I (2021) *J Mol Liq* 322:114843. <https://doi.org/10.1016/j.molliq.2020.114843>
10. Sharma P, Ramesh K, Parameshwaran R, Deshmukh SS (2022) *Case Stud Therm Eng* 30:101658. <https://doi.org/10.1016/j.csite.2021.101658>
11. Cui W, Cao Z, Li X, Lu L, Ma T, Wang Q (2021) *Powder Technol* 398:117078. <https://doi.org/10.1016/j.powtec.2021.117078>
12. Novoselska N, Rasulev B, Gajewicz A, Kuz'min V, Puzyn T, Leszczynski J (2014) *Nanoscale* 6:13986–13993. <https://doi.org/10.1039/C4NR03487B>
13. Jafari K, Fatemi MH (2020) *J Therm Anal Calorim* 142(3):1335–1344. <https://doi.org/10.1007/s10973-019-09215-3>
14. Jafari K, Fatemi MH (2020) *Adv Powder Technol* 31:3018–3027. <https://doi.org/10.1016/j.apt.2020.05.023>
15. Toropova AP, Toropov AA (2019) *J Mol Struct* 1182:141–149. <https://doi.org/10.1016/j.molstruc.2019.01.040>
16. Toropova AP, Toropov AA, Rallo R, Leszczynska D, Leszczynski J (2015) *Ecotoxicol Environ Saf* 112:39–45. <https://doi.org/10.1016/j.ecoenv.2014.10.003>
17. Lubinski L, Urbaszek P, Gajewicz A, Cronin MT, Enoch SJ, Madden JC, Leszczynska D, Leszczynski J, Puzyn T (2013) *SAR QSAR Environ Res* 24:995–1008. <https://doi.org/10.1080/1062936X.2013.840679>

18. Toropova AP, Toropov AA, Veselinović AM, Veselinović JB, Leszczynska D, Leszczynski J (2017) Multi-scale approaches in drug discovery, pp 191–221. <https://doi.org/10.1016/B978-0-08-101129-4.00008-4>
19. Toropova AP, Toropov AA, Benfenati E, Castiglioni S, Bagnati R, Passoni A, Zuccato E, Fanelli R (2018) *Process Saf Environ Prot* 118:227–233. <https://doi.org/10.1016/j.psep.2018.07.003>
20. Toropov AA, Sizochenko N, Toropova AP, Leszczynska D, Leszczynski J (2020) *J Mol Liq* 317:113929. <https://doi.org/10.1016/j.molliq.2020.113929>
21. Leone C, Bertuzzi EE, Toropova AP, Toropov AA, Benfenati E (2018) *Chemosphere* 210:52–56. <https://doi.org/10.1016/j.chemosphere.2018.06.161>
22. Sizochenko N, Jagiello K, Leszczynski J, Puzyn T (2015) *J Phys Chem C* 119:25542–25547. <https://doi.org/10.1021/acs.jpcc.5b05759>
23. Sizochenko N, Syzochenko M, Gajewicz A, Leszczynski J, Puzyn T (2017) *J Phys Chem C* 121:1910–1917. <https://doi.org/10.1021/acs.jpcc.6b08850>
24. Sizochenko N, Kar S, Syzochenko M, Leszczynski J (2018) *Int J Quant Struct Relat* 4:18–27. <https://doi.org/10.4018/ijqspr.2019010102>
25. Toropov A, Sizochenko N, Toropova A, Leszczynski J (2018) *Nanomaterials* 8:243. <https://doi.org/10.3390/nano8040243>
26. Jafari K, Fatemi MH, Toropova AP, Toropov AA (2022) *Chemom Intell Lab Syst* 222:104500. <https://doi.org/10.1016/j.chemolab.2022.104500>
27. Pordanjani AH, Aghakhani S, Afrand M, Sharifpur M, Meyer JP, Xu H, Ali HM, Karimi N, Cheraghian G (2021) *J Clean Prod* 320:128573. <https://doi.org/10.1016/j.jclepro.2021.128573>
28. Jafari K, Fatemi MH, Toropova AP, Toropov AA (2020) *Chem Phys Lett* 754:137614. <https://doi.org/10.1016/j.cplett.2020.137614>
29. Pak BC, Cho YI (1998) *Exp Heat Transf A J Therm Energy Gener Transp Stor Convers* 11:151–170. <https://doi.org/10.1080/08916159808946559>

Part VII
Possible Ways of QSPR/QSAR Evolution
in the Future

Chapter 16

On Complementary Approaches of Assessing the Predictive Potential of QSPR/QSAR Models



Andrey A. Toropov, Alla P. Toropova, Danuta Leszczynska,
and Jerzy Leszczynski

Abstract This chapter covers an overview of recent studies performed to improve the statistical tools to assess and compare different QSPR/QSAR models. The critical analysis of existing approaches to assess the predictive potential is briefly presented. The disadvantages of the systems of self-consistent models are also discussed. The potential advantages of the systems of self-consistent models are defined. A series of successful applications of the approach for several endpoints are discussed in order to confirm the potential of the approach as a tool to validate QSAR models.

Keywords QSPR/QSAR · Index ideality of correlation (IIC) · Correlation intensity index (CII) · Self-consistent models · Monte Carlo method

Abbreviation

CCC	Concordance correlation coefficient
CII	Correlation Intensity Index
IIC	Index Ideality of Correlation
MAE	Mean absolute error

A. A. Toropov (✉) · A. P. Toropova
Laboratory of Environmental Chemistry and Toxicology, Department of Environmental Health Science, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Via Mario Negri 2, 20156 Milano, Italy
e-mail: andrey.toropov@marionegri.it

A. P. Toropova
e-mail: alla.toropova@marionegri.it

D. Leszczynska
Department of Civil and Environmental Engineering, Interdisciplinary Nanotoxicity Center, Jackson State University, 1325 Lynch Street, Jackson, MS 39217-0510, USA

J. Leszczynski
Department of Chemistry, Physics and Atmospheric Sciences, Interdisciplinary Nanotoxicity Center, Jackson, MS 39217, USA

MLR	Multiple Regression Analysis
PLS	Partial Least-Squares regression analysis
RF	Random forest
RMSE	Root mean squared error
R^2	Determination coefficient
Q^2	The leave-one-out cross-validation R^2
QSPR	Quantitative structure–property relationships
QSAR	Quantitative structure–activity relationships
SMILES	Simplified molecular-input line-entry system

16.1 Introduction

Complex biochemical interactions, that define different kinds of biological activities, could be expressed as a “mathematical function” not only of the molecular structure but also some additional circumstances, such as physicochemical conditions, interactions via energy, and information effects between a substance and organisms, organs, or cells. These circumstances lead to the great complexity of prediction for biochemical endpoints since all “details” of corresponding phenomena are practically unavailable for accurate registration and analysis. Researchers have no possibility to carry out an analysis of all possible ways of the biochemical interactions, which define toxicological or therapeutically attractive effects via direct experiment. Consequently, a compromise, i.e., development of predictive models describing the above phenomena, becomes necessary. However, the estimation of the predictive potential of these models remains a vital task that by now is solved only partially.

Establishing quantitative structure–property/activity relationships (QSPRs/QSARs) between a desired endpoint (property) and structural details (descriptors) of the investigated compounds is one of the key goals of computational chemistry, and possibly, one of the directions to follow for theoretical chemistry. Perhaps both research areas could benefit from such an approach. If one considers the articles of H. Wiener [1–3] published in 1947 to be the beginning of QSAR research, then it could be argued that this direction has been successfully developing for the last 75 years.

Perhaps the most important and most unsettled issue of such techniques is the trust in quality of developed models: How to establish that the model is reliable and confirm that the model is dedicated—works well for a given substance?

In addition to the established, calculated criteria, that could evaluate quality of a model, hints such as “all models are wrong, but some of them are useful” [4] or “everything should be made as simple as possible, but not simpler” [5] can be very beneficial for applications of the QSAR techniques.

Any model must be tested before it can be used to “understand” or predict new phenomena, such as the biological activity of new compounds [6]. Nevertheless,

there are no generally accepted recommendations on how such a check should be carried out in practice.

A set of guidelines for developing validation of predictive QSPR models may be the following: First randomization of the modeled property (Y-scrambling). Second, multiple leave-many-out cross-validations. Third, application of external validation that uses rational division of a dataset into training and test sets. In addition, one should also establish the domain of applicability of a model in the chemical space [7].

Nonetheless, often y-randomization is not available to a potential user of a model due to the values of all descriptors in the pool for all compounds not being published [8]. Despite widespread use, multiple leave-one-out (as well as leave-many-out) cross-validation methods are questioned [9, 10]. There is disagreement regarding the definition of the applicability domain [11, 12]. However, most researchers working in the field of QSPR/QSAR analyses recognize the expediency of external verification, that is, the evaluation of the model with molecules not used in the construction of the model [10, 13, 14]. Evidently, an appropriate software is required to create a QSPR/QSAR models.

16.2 Software for Building Up QSPR/QSAR Models

Currently, various types of software have been developed to assist in QSPR/QSAR studies. This confirms the importance and prevalence of the problem of developing computer models of the physicochemical and biomedical behavior of various substances. Examples of this kind of software can be found on the Internet. Table 16.1 contains several examples of the mentioned software.

In addition, in order to develop QSPR/QSAR models, it is necessary to rationally distribute the available data into training and validation sets. The distribution of available data for QSPR/QSAR analyses into the training and validation sets can be done in various manners [15, 16]. Such distribution surely influences the statistical quality of QSPR/QSAR models [17–19]. Nevertheless, in the modern QSPR/QSAR research, the majority of the models are based solely on single distribution of available data into the training and validation sets. According to many authors, some rational split into training and validation sets gives better statistical results than models obtained with several random splits [20]. However, the numerical experiments point out that splits, which are successful for one approach, can be unsuccessful for another [21, 22]. Therefore, it is better to consider several splits of the data into the training and validation sets [23].

Table 16.1 List of software available on the Internet suggested for development of QSPR/QSAR models

Software	Comments	Link
CODESSA	CODESSA (Comprehensive Descriptors for Structural and Statistical Analysis) PRO is a comprehensive program for developing QSAR/QSPR by integrating all necessary mathematical and computational tools to predict property values for any chemical compound with known molecular structure	http://www.codessa-pro.com/
DRAGON	Dragon 7.0 provides an improved user interface, new descriptors, and additional features such as the calculation of fingerprints and the support for disconnected structures	https://chm.kode-solutions.net/pf/dragon-7-0/
Virtual Computational Chemistry Laboratory	This site provides free online tools that can be useful in performing computational chemistry, including building and visualizing chemical structures, calculating molecular properties, and analyzing relationships between chemical structure and properties	http://www.vcclab.org/
QSAR Research Unit in Environmental Chemistry and Ecotoxicology	The development of QSAR models for predicting the environmental behavior and biological activities of chemicals of concern, such as classical organic environmental pollutants and emerging contaminants: personal care products, pharmaceuticals, and nanoparticles	https://dunant.dista.uninsubria.it/qsar/
PaDEL-Descriptor	A software to calculate molecular descriptors. The software currently calculates 1875 descriptors	http://www.yapcwsoft.com/dd/padeldescriptor/
VEGA	The software predicts integrate traditional QSPR/QSAR models and models obtained with the read across technique	https://www.vegahub.eu/

(continued)

Table 16.1 (continued)

Software	Comments	Link
CORAL	The program provides an opportunity to develop and test QSPR/QSAR models in the “structure–property/activity” paradigm, as well as models in the “structure and experimental conditions—property/activity” paradigm (through the so-called quasi-SMILES)	http://www.insilico.eu/coral
DTC-QSAR: A complete QSAR modeling package	DTC-QSAR software is a complete modeling package providing a user-friendly, easy-to-use GUI to develop regression (MLR, PLS) and classification-based (LDA and Random Forest) QSAR models. It includes two well-known variable selection techniques, i.e., genetic algorithm and best subset selection	https://dtclab.webs.com/software-tools

16.3 The Critical Analysis of Existing Approaches to Assessing the Predictive Potential

Over the years, the number of statistical characteristics aimed to measure the predictive potential of a model has gradually increased, despite the evident attractiveness of the minimum number of criteria of the predictive potential for practical applications. On the one hand, the diversity of different standards for predicting potential could be considered as a tool to improve the quality of QSPR/QSAR models. On the other hand, this situation sometimes causes uncertainty in choosing the best model. In other words, contradictions in the recommendations of various criteria force the researcher to search for truth (i.e., the best choice) in a greater maze of possibilities. Table 16.2 contains a list of widespread criteria of the predictive potential.

16.4 Convenience and Inconvenience of Correlation

In fact, a QSPR/QSAR approach provides the user with correlations between a molecular architecture-dependent physicochemical or biochemical parameter of interest and the calculated value of the aforementioned parameter through some mathematical function that uses the molecular structure and/or controlled experimental conditions data. Suppose the model value is calculated using a small number of molecular characteristics. In that case, the predicted value becomes a very attractive alternative to

Table 16.2 Collection of the most popular criteria for the predictive potential of QSPR/QSAR

The criterion of the predictive potential ^a	References
$R = \frac{n \sum xy - \sum x \sum y}{\sqrt{(n \sum x^2 - (\sum x)^2)(n \sum y^2 - (\sum y)^2)}}$	[24]
$Q^2 = 1 - \frac{\sum (y_k - \bar{y}_k)^2}{\sum (y_k - \bar{y}_k)^2}$	[25]
$Q_{F1}^2 = 1 - \frac{\left[\sum_{i=1}^{N_{EXT}} \left(\frac{y'_i - y_i}{i} \right)^2 \right] / N_{EXT}}{\left[\sum_{i=1}^{N_{EXT}} (y_i - \bar{y}_{TR})^2 \right] / N_{EXT}}$	[26]
$Q_{F2}^2 = 1 - \frac{\left[\sum_{i=1}^{N_{EXT}} \left(\frac{y'_i - y_i}{i} \right)^2 \right] / N_{EXT}}{\left[\sum_{i=1}^{N_{EXT}} (y_i - \bar{y}_{EXT})^2 \right] / N_{EXT}}$	[26]
$Q_{F3}^2 = 1 - \frac{\left[\sum_{i=1}^{N_{EXT}} \left(\frac{y'_i - y_i}{i} \right)^2 \right] / N_{EXT}}{\left[\sum_{i=1}^{N_{TR}} (y_i - \bar{y}_{TR})^2 \right] / N_{TR}}$	[26]
$\bar{R}_m^2 = \frac{R_m^2(x, y) + R_m^2(y, x)}{2}$	[27]
$\Delta R_m^2 = R_m^2(x, y) - R_m^2(y, x) $	
$CCC = \frac{2 \sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2 + n(\bar{x} - \bar{y})^2}$	[28]
$IIC_c = r_C \frac{\min(-MAE_c, +MAE_c)}{\max(-MAE_c, +MAE_c)}$	[29]
$-MAE_c = \frac{1}{-N} \sum_{k=j}^{-N} \Delta_k , \Delta_k 0; -N \text{ is the number of } \Delta_k < 0$	
$+MAE_c = \frac{1}{+N} \sum_{k=j}^{+N} \Delta_k , \Delta_k 0; +N \text{ is the number of } \Delta_k \geq 0,$	
$\Delta_k = \text{observed}_k - \text{calculated}_k$	
$CII_c = 1 - \sum (\Delta R_j^2 > 0)$	[30]
$\Delta R_j^2 = R_j^2 - R^2$	

^a x and y are experimental and predicted values of endpoint; n is the number of compounds in a set; R is the Pearson correlation coefficient; Q^2 is cross-validated R^2 ; CCC is concordance correlation coefficient; IIC_c is the index of ideality of correlation; MAE is mean absolute error; CII_c is the correlation intensity index

direct experimental determination, which requires time for the experiment, reagents, and a certain level of personnel qualification. It is to be noted that if such value is calculated using larger number of characteristics it should be even more useful since for larger pool of parameters more experiments are required. Interestingly, such developed correlation may lead to hypotheses about the respective mechanisms of molecular action. However, the legitimacy of these hypotheses cannot be tested without a statistically significant number of tests of the validity of this correlation

for substances at least theoretically suitable for their respective applications. In principle, this situation can be considered quite acceptable for solving local problems such as the choice of economically and environmentally acceptable dyes, packaging materials, thermal insulators, and others. At the same time, such a situation is unacceptable or even risky when designing drugs. One has to remember that a correlation isolated from reality may become the basis for many erroneous assumptions and interpretations.

Without doubt debatable issues related to QSAR include an erroneous association of correlation with causation. In addition, it is important to note that the predictive potential of a model is not necessarily a measure of its utility [31].

Developed correlation often has a very convenient form [32]: without completely solving a scientific or technical problem. The correlations developed for QSAR allow, at the cost of relatively small expenses, to outline prospective targets for in-depth scientific research. Thus, paraphrasing a famous aphorism, “correlation often has the first word, but never the last.”

16.5 Convenience and Inconvenience of Causation

If causation is better than correlation, the question arises: should one use it or at least try everything necessary to turn correlations into causation? First of all, it should be borne in mind that many problems and situations exclude the possibility for devoting a reasonable amount of time to move from correlations to causation.

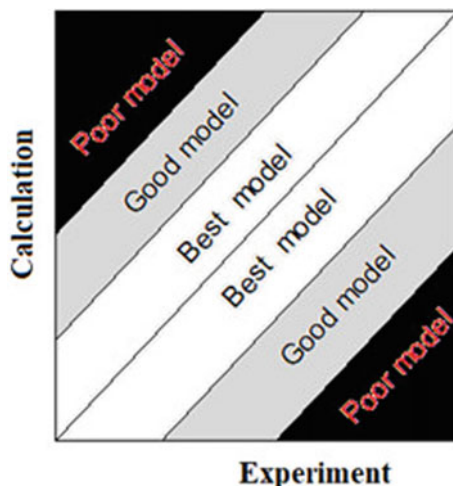
Causality is not eternal and is not always reliable, and the problem arises with the need to monitor whether the reason to use it has lost its relevance.

Let's explain this point in more details. Initially, the study and assessment of odor was based on intensity (causality is defined as a strong odor produces a strong effect). However, subsequently such assessment moved to a valuation based on the intensity of effects on receptors (now causality is defined as a great change of receptors that has a strong effect) [33]. Another example: advertising the availability and reliability of cars can lead to a decrease in demand [34].

Therefore, causality cannot be measured. The strange idea that the predictive potential is not actually a measure of the quality of the model [31] can be transformed into the opinion that “causality cannot be proven.” Causality can be confirmed in 100 experiments but failed in the experiment #101.

Thus, causality does not have its own measure of reliability, unlike correlation, which can be measured via calculated criteria.

Fig. 16.1 Portraits of good and poor QSPR/QSAR models



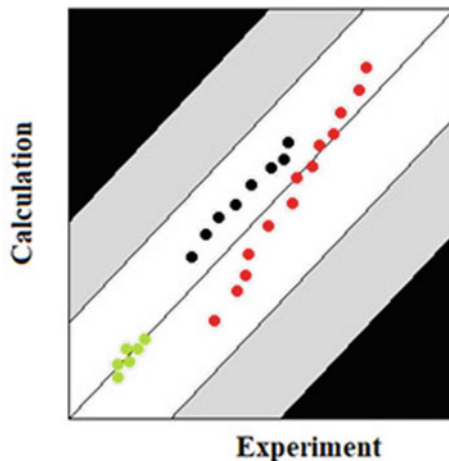
16.6 Note on “Secrets of QSPR/QSAR”

It is customary to characterize the quality of the model by the values of the coefficient of determination (the square of the correlation coefficient) and the standard deviation or the value of the average absolute error. However, it is usually not mentioned that the main advantage of the QSPR/QSAR model is how the points are located in the coordinates of "experiment vs. calculation" relative to the diagonal of some square. Figure 16.1 shows zones one can establish to detect the quality of models in the first approximation. The simplest representation of correlations is presented. It is both correct and uninformative. Nevertheless, this is the basis for the construction of QSPR/QSAR. Figure 16.2 confirms that defining a “good model” is not straightforward.

16.7 Index Ideality of Correlation (*IIC*)

The fuzziness of the manifestations of the actual world leads to the fact that mathematical idealizations cannot reliably represent a significant part of the real world. There is no easy way to define the term “suitable unambiguously.” For example, how to select a suitable car? In some cases, a small and compact car is sufficient. In other cases, a powerful heavy truck may be more ideal than a nimble supercar and vice versa. Clearly, the “class of all real numbers which are much greater than 1,” or “the class of beautiful women,” or “the class of tall men,” do not constitute classes or sets in the usual mathematical sense of these terms. Yet, the fact remains that such imprecisely defined “classes” play an important role in human thinking, particularly in pattern recognition and information communication. Essentially, a fuzzy set is

Fig. 16.2 Portraits of poor QSPR/QSAR models are placed in a place where the best models are expected (green dots describe poor correlation which can be detected in other scale; red dots display situation where slope of a regression model is not suitable for the validation set; black dots refer to a case where intercept of a regression model is not suitable for the validation set)



a natural way of dealing with problems in which the source of imprecision is the absence of sharply defined criteria of class membership [35].

One of the quite pressing examples of fuzziness is the correlations observed for QSPR/QSAR models. The fuzzy correlations that are “fuzzy” consequently to the value of the coefficient of determination can be ideal non-linear correlations, the graphical appearance of which will prompt the user that he is dealing with an “explicit” correlation.

Obviously, without a graphical image, the user, focusing on the coefficient of determination only, is unlikely to be able to assess the quality of a clear non-linear correlation.

At the same time, situations often arise when comparing large sets of models without visualizing them is necessary. This makes it desirable to develop efficient alternatives to the traditional coefficient of determination used in numerous statistical investigations.

The idealization (or simplification) is one of the most common approaches to study complex phenomena in the field of natural sciences, e.g., ideal gas, ideal solution, ideal crystals, and ideal symmetry [29].

The index of ideality of correlation (IIC_C) is one of the possibilities to evaluate the quality of models in the above sense. It is well known that in practice, it is necessary to consider large numbers of pair's correlations for quantities where there is no data on the standard error of estimation or the mean absolute error. For such situations, the index of ideality of correlation is not suitable.

The IIC_C is calculated with data on the calibration set as the following:

$$IIC_C = r_c \frac{\min(-MAE_c, +MAE_c)}{\max(-MAE_c, +MAE_c)} \quad (16.1)$$

$$\min(x, y) = \begin{cases} x, & \text{if } x < y \\ y, & \text{otherwise} \end{cases} \quad (16.2)$$

$$\max(x, y) = \begin{cases} x, & \text{if } x > y \\ y, & \text{otherwise} \end{cases} \quad (16.3)$$

$${}^{-}\text{MAE}_c = \frac{1}{-N} \sum |\Delta_k|, \quad {}^{-}N \text{ is the number of } \Delta_k < 0 \quad (16.4)$$

$${}^{+}\text{MAE}_c = \frac{1}{+N} \sum |\Delta_k|, \quad {}^{+}N \text{ is the number of } \Delta_k \geq 0 \quad (16.5)$$

$$\Delta_k = \text{observed}_k - \text{calculated}_k \quad (16.6)$$

The observed and calculated are corresponding values of an endpoint.

The index of ideality of correlation improves models' predictive potential based on so-called optimal descriptors calculated with SMILES [29]. Thus, it is advised to apply the index for QSPR/QSAR analyses of different endpoints.

16.8 Correlation Intensity Index (*CII*)

The fact that the index of ideality of correlation cannot be used in situations where there is no data on *RMSE* or *MAE* forces one to look for some alternative, that is, an index that makes it possible to quickly assess the quality of a "hidden" but promising correlation. The correlation intensity index is an attempt to develop a statistical index that could play the role of detector of "attractive hidden" correlations [30]. The application of *CII* in QSPR/QSAR analyses indicates that the contribution of the *CII* improves the predictive potential of QSPR/QSAR models based on optimal descriptors calculated with SMILES [30, 36–42]. These models can be applied to systematize knowledge in various areas including physical chemistry, biochemistry, ecology, and medical sciences [30, 36–42].

The *CII_c* is calculated as follows:

$$\text{CII}_c = 1 - \sum \text{Protest}_k \quad (16.7)$$

$$\text{Protest}_k = \begin{cases} R_k^2 - R^2, & \text{if } R_k^2 - R^2 > 0 \\ 0, & \text{otherwise} \end{cases} \quad (16.8)$$

The R^2 is the correlation coefficient for a set that contains n substances. The R_k^2 is the correlation coefficient for $n - 1$ substances of a set after removing of k -th substance. Hence, if the $(R_k^2 - R^2)$ is larger than zero, the k -th substance is an "opponent" for the correlation between experimental and predicted values of the set. A small sum of "protests" means a more "intensive" correlation.

16.9 Can *IIC* and *CII* Be Useful?

IIC and *CII* are suitable for improving the optimal descriptors calculated with SMILES, but can they be applied similarly to the traditional correlation coefficient or the concordance correlation coefficient?

Figure 16.3 contains a collection of correlations and non-correlation that differ in nature. In other words, the mentioned collection contains linear and non-linear correlations with different levels of fuzziness.

The cases marked *a*, *b*, and *c* represent not-linear correlations. The *IIC* can indicate a good correlation if its value is close to 1. Hence, *IIC* does classify such cases as non-correlations. *CII* also is able to demonstrate a good correlation if its value is close to 1. Hence, *CII* does classify these cases as correlations. Therefore, *IIC* and *CII* contradict each other in the situations denoted by *a*, *b*, and *c*.

The situations marked *d*, *e*, and *f* represent linear correlations of varying degrees of fuzziness. *IIC* detect a good correlation for *d* and *f*, but a poor correlation for *e*. The case *e* is characterized by the location of all dots far from the diagonal. Regarding the fuzziness of the situation, *e* and *f* are identical.

Cases *g*, *h*, and *i* represent linear correlations of varying degrees of fuzziness. The values of *CII* in all cases "recognize the correlation", while *IIC* (except in the case of *g*), rejects them.

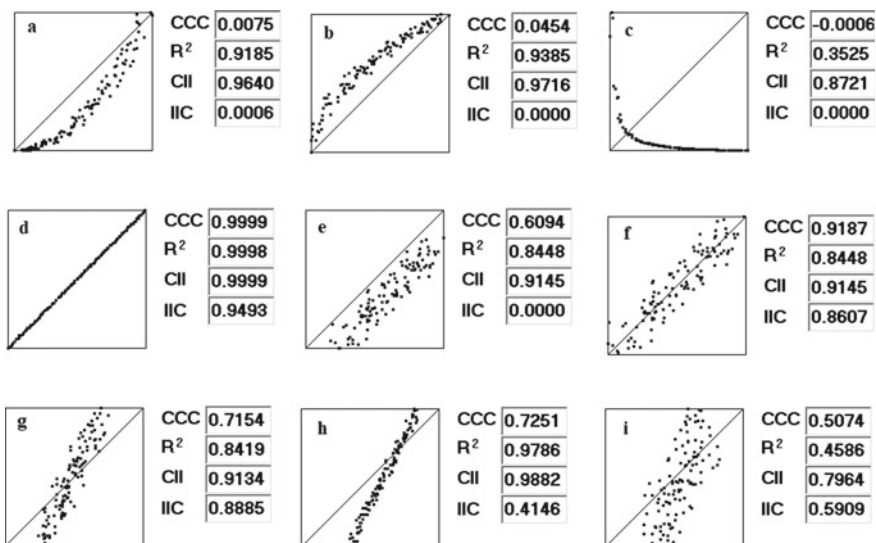


Fig. 16.3 Estimation of different correlations via different criteria of the predictive potential. *CCC* = concordance correlation coefficient; *R*² = determination coefficient; *CII* = correlation intensity index; *IIC* = index of ideality of correlation

Summing up, *IIC* to some extent, it is similar to *CCC*, but its assessment is more complex than the assessment of *CCC*. At the same time, *CII* is somewhat similar to R^2 , but its review is softer than the assessment of R^2 .

In the process of developing and using the CORAL software (<http://www.insilico.eu/coral>), a number of experiments were carried out. The following important questions have been asked during these studies:

1. Is it possible to obtain correlations suitable for prediction of various physico-chemical and/or biochemical characteristics of substances based on the correlation weighting of molecular features extracted from SMILES?
2. Is it possible to improve the predictive potential of such models using *IIC*?
3. Is it possible to improve the predictive potential of such models using *CII*?

Below we address all the questions.

16.10 Is It Possible to Improve the Predictive Potential of Such Models Using *IIC*?

Figure 16.4 shows the impact of using the *IIC* on the optimization process. A record of the computational process without the *IIC* shows a gradual increase in the correlation coefficient for the training set, which is first accompanied by an increase in the correlation coefficient for the calibration and test sets. However, for the latter, this growth reaches a maximum and then gradually decreases.

A record of optimization using *IIC* is quite different. For both training sets (active and passive) and control sets (calibration set and validation set), the correlation

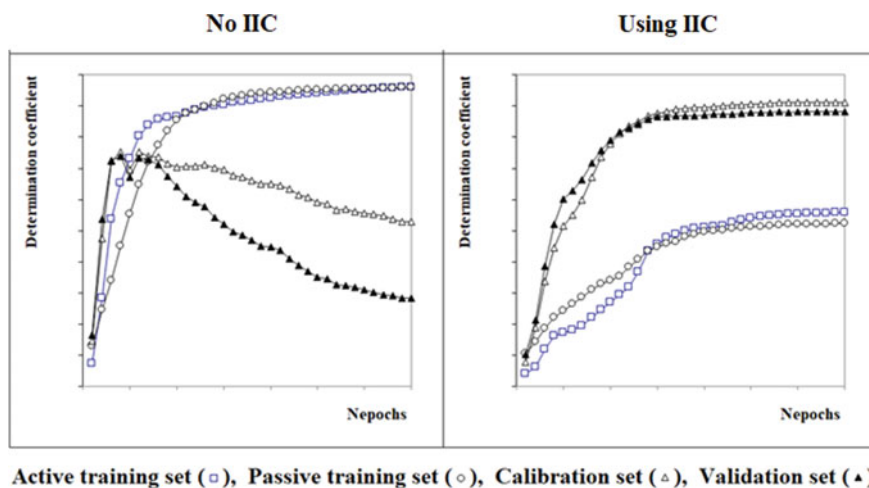


Fig. 16.4 Record of the Monte Carlo optimization with and without *IIC*

coefficient gradually increases. However, in the case of optimization using *IIC*, the statistical quality of the model for training samples is noticeably more modest. In other words, *IIC* contribution improves correlations for validation sets (including the external validation set) but to the detriment of training sets (i.e., the statistical characteristics for the active and passive training sets are reducing).

Table 16.3 contains a collection of computer experiments described in the literature to test the *IIC*. It can be seen (Table 16.3) that *IIC* contributes to improve the predictive potential of the proposed models of various types of biological activity.

Table 16.3 Applying the Monte Carlo method to build up QSPR/QSAR models using *IIC*

Endpoint	The statistical quality	Comments	References
Cytotoxicity of 2-phenylindole derivatives against breast cancer cells	Without <i>IIC</i> Active training set $R^2 = 0.8435$, $MAE = 0.376$; validation set $R^2 = 0.9017$, $MAE = 0.211$ Using <i>IIC</i> Active training set $R^2 = 0.8037$, $MAE = 0.452$; validation set $R^2 = 0.9685$, $MAE = 0.128$	Three random splits confirm the predictive potential of the approach	[43]
The experimental values measured for EC_{50} (effective molar concentration) (mol/L) are represented by negative decimal logarithm pEC_{50}	Without <i>IIC</i> Active training set $R^2 = 0.8921$, $RMSE = 0.291$; validation set $R^2 = 0.9062$, $RMSE = 0.267$ using <i>IIC</i> Active training set $R^2 = 0.7877$, $RMSE = 0.409$; validation set $R^2 = 0.9515$, $RMSE = 0.223$	Three random splits confirm the predictive potential of the approach	[44]
The sweetness potential ($\log Sw$) is represented by the ratio of the concentration of the test compound in water with an equivalent concentration of sucrose in water	Without <i>IIC</i> Active training set $R^2 = 0.8921$, $RMSE = 0.291$; validation set $R^2 = 0.9062$, $RMSE = 0.267$ using <i>IIC</i> Active training set $R^2 = 0.7877$, $RMSE = 0.409$; validation set $R^2 = 0.9515$, $RMSE = 0.223$	Three random splits confirm the predictive potential of the approach	[45]

(continued)

Table 16.3 (continued)

Endpoint	The statistical quality	Comments	References
Cell viability (%) for human breast cancer cell line MCF-7	Without <i>IIC</i> Active training set $R^2 = 0.9399$, $MAE = 6.0$; validation set $R^2 = 0.9272$, $MAE = 6.1$ Using <i>IIC</i> Active training set $R^2 = 0.9172$, $MAE = 7.1$; validation set $R^2 = 0.9416$, $MAE = 7.1$	Three random splits confirm the predictive potential of the approach	[46]
Glucokinase activators; experimental values were changed into negative decimal logarithm (pEC_{50})	Without <i>IIC</i> Calibration set $R^2 = 0.2635$, validation set $R^2 = 0.7209$ Using <i>IIC</i> Calibration set $R^2 = 0.7190$, validation set $R^2 = 0.7936$	Statistics for calibration set and validation set only	[47]
Enhancement of azo dye adsorption affinity for cellulose fiber	Without <i>IIC</i> Training set $R^2 = 0.9972$, $RMSE = 0.256$; validation set $R^2 = 0.7597$, $RMSE = 2.072$ Using <i>IIC</i> Training set $R^2 = 0.7190$, $RMSE = 1.43$; validation set $R^2 = 0.7936$, $RMSE = 1.200$	Three random splits confirm the predictive potential of the approach	[48]

16.11 Is It Possible to Improve the Predictive Potential of Such Models Using *CII*?

The correlation intensity index was proposed somewhat later than the correlation ideality index. Therefore, *CII* has been studied to a lesser extent than *IIC*. However, some results point to this index's ability to measure the predictive power of models. Table 16.4 contains some examples of the use of *CII* in building up models for different endpoints.

16.12 Testing Assumptions About the Significance of *IIC* and *CII*

It can be assumed that the Monte Carlo optimization carried out taking into account the contributions of the *IIC* and *CII* will give better models than such an optimization

Table 16.4 Applying the Monte Carlo method to build up QSPR/QSAR models using *CII*

Endpoint	The statistical quality	Comments	References
Zeta potentials (ζ) in metal oxide nanoparticles	No <i>IIC</i> , No <i>CII</i> $\overline{R}_v^2 = 0.7012$ Use <i>IIC</i> (only) $\overline{R}_v^2 = 0.7590$ Use <i>IIC</i> and <i>CII</i> $\overline{R}_v^2 = 0.8674$	The use of <i>IIC</i> improved the predictive power, but the combined use of <i>IIC</i> and <i>CII</i> further increased the model's predictive power	[38]
Biological activity of anti-influenza single-stranded DNA aptamers	No <i>IIC</i> , No <i>CII</i> $\overline{R}_v^2 = 0.7501$ Use of <i>IIC</i> (only) $\overline{R}_v^2 = 0.7687$ Use <i>IIC</i> and <i>CII</i> $\overline{R}_v^2 = 0.8801$	The use of <i>IIC</i> improved the predictive power, but the use of <i>CII</i> further increased the model's predictive power	[40]
Skin sensitivity (<i>PEC3</i>)	No <i>IIC</i> , No <i>CII</i> $\overline{R}_v^2 = 0.672$ Use of <i>IIC</i> $\overline{R}_v^2 = 0.726$ Use of <i>CII</i> $\overline{R}_v^2 = 0.744$ Use of <i>IIC</i> and <i>CII</i> $\overline{R}_v^2 = 0.779$	The use of <i>IIC</i> improved the predictive power, but the combined use of <i>IIC</i> and <i>CII</i> further increased the model's predictive power	[49]

based on only the *IIC*, or only the *CII*. To test this assumption, some data must be used. In particular, the toxicity data discussed in [50] provide a pool of data that could be applied for this purpose. For the specified check, the following descriptors were used, calculated on the basis of SMILES:

$$DCW(T^*, N^*) = \sum CW(APP_k) + \sum CW(S_k) + \sum CW(SS_k) + \sum CW(SSS_k) \quad (16.9)$$

The Monte Carlo optimization is a tool to calculate correlation weights for the descriptor. Two target functions, TF_1 and TF_2 , for the Monte Carlo optimization should be examined.

$$TF_0 = r_{AT} + r_{PT} - |r_{AT} - r_{PT}| \times 0.1 \quad (16.10)$$

$$TF_1 = TF_0 + IIC_c \times 0.5 \quad (16.11)$$

$$TF_2 = TF_1 + CII_c \times 0.5 \quad (16.12)$$

Table 16.5 contains the statistical characteristics of models studied in the case of target functions TF_1 and TF_2 . One can see the first model calculated with *IIC* is

better than the model suggested in the literature [50]. Interestingly, the second model studied in the case of involving both *IIC* and *CII* is better than model obtained using *IIC* solely without *CII* contribution.

Table 16.5 Statistical characteristics of models calculated using *IIC* and *CII*

Set	n	R ²	CCC	<i>IIC</i>	<i>CII</i>	Q ²	RMSE	F
<i>Split 1</i>								
Active training	97	0.5571	0.7156	0.6459	0.8197	0.5371	1.04	120
Passive training	102	0.5584	0.6358	0.6738	0.7450	0.5415	1.23	126
Calibration	99	0.8184	0.8923	0.9046	0.9026	0.8116	0.463	437
Validation	102	0.7907	0.8890	0.8813	0.8841		0.493	
Active training	97	0.4934	0.6607	0.6603	0.8079	0.4712	1.11	93
Passive training	102	0.5442	0.6025	0.6696	0.7502	0.5265	1.26	119
Calibration	99	0.8863	0.9240	0.9414	0.9454	0.8815	0.383	756
Validation	102	0.8439	0.9183	0.7722	0.9178		0.422	
<i>Split 2</i>								
Active training	103	0.5648	0.7219	0.6820	0.8090	0.5493	1.04	131
Passive training	101	0.5652	0.6690	0.5689	0.7698	0.5451	1.14	129
Calibration	98	0.7259	0.8516	0.8520	0.8634	0.7153	0.617	254
Validation	98	0.7614	0.8684	0.8548	0.8566		0.600	
Active training	103	0.4673	0.6370	0.6203	0.8128	0.4503	1.16	89
Passive training	101	0.5736	0.5959	0.6572	0.7843	0.5560	1.16	133
Calibration	98	0.8165	0.8912	0.9034	0.9162	0.8078	0.492	427
Validation	98	0.7844	0.8718	0.6599	0.8910		0.549	
<i>Split 3</i>								
Active training	103	0.5815	0.7354	0.7194	0.8040	0.5654	0.987	140
Passive training	103	0.6149	0.6260	0.6148	0.7987	0.5979	1.25	161
Calibration	96	0.7204	0.8473	0.8488	0.8552	0.7095	0.586	242
Validation	98	0.7884	0.8750	0.7113	0.8686		0.645	
Active training	103	0.5181	0.6826	0.6531	0.8057	0.4995	1.06	109
Passive training	103	0.5920	0.5965	0.6307	0.7980	0.5757	1.27	147
Calibration	96	0.8395	0.9134	0.9158	0.9212	0.8323	0.413	492
Validation	98	0.8084	0.8982	0.7760	0.8788		0.528	

16.13 The Comparison of Criteria of the Predictive Potential of QSPR/QSAR

QSPR/QSAR is the applicative theoretical tool of modern natural sciences. This approach provides qualitative (yes/no) and/or quantitative (how much) models for attribute of various substances. What does it take to recognize "the model is usable"? Suppose there are some experimental and model (calculated) values of a physico-chemical property or biological activity. Those experimental and computed value pairs in the external test set are close to the diagonal of the experiment versus calculations relationship. In that case, there are good reasons to say, "The model can be recommended for practical application" (Fig. 16.5). Vice versa, if these values show large dispersion (relatively to the diagonal), the model is rather poor than good.

However, in practice, the model developer, at best, receives information about the mentioned triangle only after building up the model. Therefore, mathematical criteria to assess the statistical quality of the model are essential for the practical development of models. There are many such criteria. Nevertheless, unfortunately, in practice, these criteria often do not guarantee that the model is suitable for use. Naturally, under such circumstances, new statistical criteria are searched for, as well as algorithms designed to solve the problem of a reliable assessment of the predictive potential of models. There are two counter-trends in developing quantitative and qualitative models of physicochemical properties and biological activity. Models must be deterministic, that is, they must predict the behavior of substances well, even though the sets of substances for which the forecast is required are immensely large and random (not deterministic). Under such conditions, a compromise that satisfies all interested parties becomes unlikely (or even non-possible at all), but

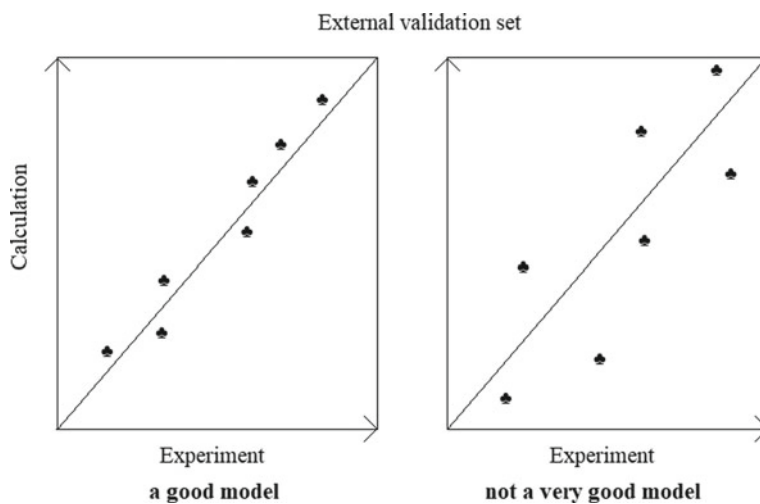


Fig. 16.5 Comparison of predictive potential of two models

developments in this area are unthinkable without some kind of compromise. Table 16.6 contains a collection of statistical criteria for evaluating the predictive potential for various QSPR/QSAR approaches [24–31]. Unfortunately, "the unreliability of the reliability criteria" is an unpleasant but reliable rule [49]. Nevertheless, since these criteria are necessary and, in some cases, useful, experiments designed to compare their reliability (unreliability) may be considered as quite appropriate and even useful.

Table 16.6 contains the results of a comparison of fifteen models according to the rating defined as

$$\text{Rating} = \begin{cases} \text{Correct if } CR1_c > CR2_c \text{ and } CR1_v > CR2_v \\ \text{or if } CR1_c < CR2_c \text{ and } CR1_v < CR2_v \\ \text{Wrong} & \text{Otherwise} \end{cases} \quad (16.13)$$

Table 16.6 contains 105 comparisons of models. The displayed data indicates that all statistical criteria (Table 16.2) give reasonable good assess for predictive potential, but no criteria avoids mistakes (wrong assessment).

16.14 The System of Self-consistent Models

A rather attractive alternative to using potential predictive criteria is constructing so-called systems of self-consistent models. The scheme for creating the system of self-consistent models is as follows:

Each i -th model has an i th validation set. The validation sets must be non-identical. It is important to determine whether the arbitrary model can be used for a random validation set. If the answer is yes, these different models should be considered self-consistent ones.

The measure of self-consistency is the average and dispersion of the correlation coefficient on different validation sets. The matrix can represent the corresponding computational experiments:

$$\begin{bmatrix} (M_1 : V'_1 \rightarrow Rv_{11}^2) \cdots (M_n : V'_1 \rightarrow Rv_{n1}^2) \\ \vdots \\ (M_1 : V'_n \rightarrow Rv_{n5}^2) \cdots (M_n : V'_n \rightarrow Rv_{nn}^2) \end{bmatrix} \quad (16.14)$$

the M_i is an i -th model; the V_j is the list of compounds applied as the validation set in the case of j -th split; the Rv_{ij}^2 is the correlation coefficient observed for the j -th validation set if applied i -th model. The n is the total number of models (splits). Currently, the system of n models vs n splits was examined. However, it should be noted that the number of models, and the number of splits can be different.

Table 16.6 Comparison of fifteen models

Models	R^2	CCC	IIC	CII	Q^2	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	$\langle R^2_m \rangle$
[1better 2]	C	C	C	C	C	C	C	C	C
[1better 3]	C	C	C	C	C	W	W	C	C
[1poorer 4]	W	W	W	W	W	W	W	W	W
[1poorer 5]	W	W	W	W	W	W	W	W	W
[1poorer 6]	W	W	W	W	W	C	W	C	W
[1poorer 7]	W	W	W	W	W	W	W	W	W
[1better 8]	C	C	C	C	C	C	C	C	C
[1better 9]	C	C	C	C	C	C	C	C	C
[1better 10]	C	C	C	C	C	C	C	W	C
[1better 11]	C	C	C	C	C	C	C	C	C
[1better 12]	C	C	C	C	C	C	C	C	C
[1better 13]	C	C	C	C	C	W	C	W	C
[1poorer 14]	W	W	W	W	W	W	W	W	W
[1better 15]	C	C	C	C	C	W	C	W	C
[2better 3]	C	C	C	C	C	C	C	C	C
[2better 4]	C	C	C	C	C	C	C	C	C
[2better 5]	C	C	C	C	C	C	C	C	C
[2better 6]	C	C	C	W	C	C	C	C	C
[2better 7]	C	C	C	C	C	C	C	C	C
[2better 8]	C	C	C	C	C	W	C	W	W
[2poorer 9]	C	C	C	C	C	C	C	C	C
[2poorer 10]	C	C	C	C	C	C	C	C	C
[2poorer 11]	C	C	C	W	C	C	C	W	C
[2poorer 12]	W	W	W	W	W	W	W	W	W
[2poorer 13]	C	C	C	C	C	C	C	C	C
[2poorer 14]	C	C	C	C	C	C	C	C	C
[2poorer 15]	C	C	C	C	C	C	C	C	C
[3poorer 4]	W	W	W	W	W	W	W	W	C
[3poorer 5]	W	W	W	W	W	W	W	W	W
[3poorer 6]	W	W	W	W	W	W	W	C	W
[3poorer 7]	C	W	C	C	C	W	W	W	W
[3better 8]	C	C	C	C	C	C	C	C	C
[3better 9]	C	C	C	C	C	C	C	W	C
[3poorer 10]	W	W	W	W	W	W	W	C	W
[3better 11]	C	C	C	C	C	C	C	C	C
[3better 12]	C	C	C	C	C	C	C	C	C

(continued)

Table 16.6 (continued)

Models	R^2	CCC	IIC	CIH	Q^2	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	$\langle R^2_m \rangle$
[3better 13]	C	C	C	C	C	W	C	W	C
[3poorer 14]	C	W	C	C	C	W	W	W	W
[3better 15]	C	C	C	C	C	C	C	W	C
[4poorer 5]	C	C	C	C	C	W	C	C	W
[4better 6]	C	C	C	C	C	W	C	W	C
[4poorer 7]	C	W	C	C	C	C	C	C	W
[4better 8]	C	C	C	C	C	C	C	C	C
[4better 9]	W	C	W	W	W	W	C	W	C
[4better 10]	C	C	C	C	C	W	C	W	C
[4better 11]	C	C	C	C	C	C	C	C	C
[4better 12]	C	C	C	W	C	C	C	W	C
[4better 13]	W	C	W	W	W	W	C	W	C
[4poorer 14]	C	W	C	C	C	C	C	C	W
[4better 15]	C	C	C	C	C	W	C	W	C
[5better 6]	C	C	C	C	C	W	C	W	C
[5better 7]	C	W	C	C	C	C	W	W	W
[5better 8]	C	C	C	C	C	C	C	C	C
[5better 9]	W	C	W	W	W	W	C	W	W
[5better 10]	C	C	C	C	C	W	C	W	C
[5better 11]	C	C	C	C	C	C	C	C	C
[5better 12]	C	C	C	W	C	C	C	C	C
[5better 13]	C	C	C	C	C	W	C	W	W
[5better 14]	C	W	C	C	C	C	W	W	W
[5better 15]	C	C	C	C	C	W	C	W	C
[6poorer 7]	C	C	C	C	C	W	C	W	W
[6better 8]	C	C	C	C	C	C	C	C	C
[6better 9]	W	W	W	W	W	C	W	C	W
[6better 10]	W	W	W	W	W	C	W	C	W
[6better 11]	C	C	C	C	C	C	C	C	W
[6better 12]	W	W	W	W	W	C	W	C	W
[6better 13]	W	W	W	W	W	W	W	W	W
[6poorer 14]	C	C	C	C	C	W	C	W	W
[6better 15]	W	W	W	W	W	C	W	C	C
[7better 8]	C	C	C	C	C	C	C	C	W
[7better 9]	C	W	C	C	C	W	C	W	W
[7better 10]	C	W	C	C	C	C	C	W	W

(continued)

Table 16.6 (continued)

Models	R^2	CCC	IIC	CIH	Q^2	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	(R^2_m)
[7better 11]	C	C	C	C	C	C	C	C	W
[7better 12]	C	W	C	C	C	C	C	C	W
[7better 13]	C	W	C	C	C	W	C	W	W
[7better 14]	C	C	C	C	C	C	C	C	C
[7better 15]	C	C	C	C	C	W	C	W	W
[8poorer 9]	C	C	C	C	C	C	C	C	C
[8poorer 10]	C	C	C	C	C	C	C	C	C
[8poorer 11]	C	C	C	C	C	W	C	W	C
[8poorer 12]	C	C	C	C	C	C	C	C	C
[8poorer 13]	C	C	C	C	C	C	C	C	C
[8poorer 14]	C	C	C	C	C	C	C	C	W
[8poorer 15]	C	C	C	C	C	C	C	C	C
[9better 10]	W	W	W	W	W	W	C	C	W
[9better 11]	C	C	C	C	C	C	C	C	C
[9better 12]	C	W	C	W	C	C	C	C	C
[9better 13]	C	C	C	C	C	W	C	W	C
[9better 14]	C	W	C	C	C	W	C	W	W
[9better 15]	C	C	C	C	C	W	C	W	C
[10better 11]	C	C	C	C	C	C	C	C	C
[10better 12]	C	W	C	W	C	C	C	C	W
[10better 13]	W	C	W	W	W	W	C	W	W
[10poorer 14]	C	W	C	C	C	C	C	W	W
[10better 15]	C	C	C	C	C	W	C	W	C
[11poorer 12]	W	W	W	W	W	W	W	W	W
[11poorer 13]	C	C	C	C	C	C	C	C	C
[12poorer 14]	C	C	C	C	C	C	C	C	W
[11poorer 15]	C	C	C	C	C	C	C	C	W
[12poorer 13]	C	W	C	W	C	C	C	C	C
[12poorer 14]	C	W	C	C	C	C	C	C	W
[12poorer 15]	W	W	W	W	W	C	W	C	W
[13poorer 14]	C	W	C	C	C	W	C	W	W
[13poorer 15]	W	W	W	W	W	W	W	W	W
[14better 15]	C	C	C	C	C	W	C	W	W
Correct frequency	82	69	82	75	82	61	82	57	58
Correct percentage	0.78	0.66	0.78	0.71	0.78	0.58	0.78	0.54	0.55

The X better Y means that the statistical quality of X-th model is better; X poorer Y means that the statistical quality of Y-th model is better; C = correct, W = wrong

16.14.1 *Examples of Successful Applications of Self-consistent Models*

Systems of self-consistent models as a method of development and, most importantly, tests of the statistical quality of models have passed the first approbation, confirming certain advantages of this methodology. The methodology was used to development of models of the biological activity of nanoparticles having the same nanocore but different surface modifiers (small organic molecules) [51] as well as for building up models for the octanol/water partition coefficient of gold nanoparticles [52]. In addition, the approach gives reasonably well models applicable for the discovery of antiviral drugs [53]. The method provides the statistically reasonable recommendations to select agents in Alzheimer's disease treatment [54]. Finally, the systems of self-consistent models offers promising results related to modeling vapor pressure [55] and the physicochemical behavior of polymers [56].

16.15 Conclusions

Reliable verification of the predictive potential of developed models is an ideal but perhaps unattainable task of modern natural sciences. This obviously also apply to the QSAR/QSPR studies. There are several predictive potential criteria (Table 16.2). Each of these criteria can assess the predictive potential of models for external sets of substances included in the domain of applicability of the model. However, all these criteria represent values requiring additional checks. The development of systems of self-consistent models is an alternative to evaluating models through predictive potential criteria. In fact, this is an algorithm for checking the quality of the model based on considering QSPR/QSAR as random events occurring from random splits into training and testing sets.

References

1. Wiener H (1947) *J Am Chem Soc* 69(1):17–20. <https://doi.org/10.1021/ja01193a005>
2. Wiener H (1947) *J Chem Phys* 15(10):766. <https://doi.org/10.1063/1.1746328>
3. Wiener H (1947) *J Am Chem Soc* 69(11):2636–2638. <https://doi.org/10.1021/ja01203a022>
4. Box GEP (1976) *J Am Stat Assoc* 71(356):791–799. <https://doi.org/10.1080/01621459.1976.10480949>
5. Maccaferri G, Lacaille J-C (2003) *Trends Neurosci* 26(10):564–571. <https://doi.org/10.1016/j.tins.2003.08.002>
6. Wold S, Sjöström M, Eriksson L (2001) *Chemom Intell Lab Syst* 58(2):109–130. [https://doi.org/10.1016/S0169-7439\(01\)00155-1](https://doi.org/10.1016/S0169-7439(01)00155-1)
7. Tropsha A, Gramatica P, Gombar VK (2003) *QSAR Comb Sci* 22(1):69–77. <https://doi.org/10.1002/qsar.200390007>
8. Rücker C, Rücker G, Meringer M (2007) *J Chem Inf Model* 47(6):2345–2357. <https://doi.org/10.1021/ci700157b>

9. Golbraikh A, Tropsha A (2002) *J Mol Graph Model* 20(4):269–276. [https://doi.org/10.1016/S1093-3263\(01\)00123-1](https://doi.org/10.1016/S1093-3263(01)00123-1)
10. Majumdar S, Basak SC (2018) *Curr Comput-Aided Drug Des* 14(1):5–6. <https://doi.org/10.2174/157340991401180321112006>
11. Netzeva TI, Worth AP, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, Jaworska JS, Kahn S, Klopman G, Marchant CA, Myatt G, Nikolova-Jeliazkova N, Patlewicz GY, Perkins R, Roberts DW, Schultz TW, Stanton DT, Van De Sandt JJM, Tong W, Veith G, Yang C (2005) *ATLA Altern Lab Anim* 33(2):155–173. <https://doi.org/10.1177/026119290503300209>
12. Tropsha A, Golbraikh A (2007) *Curr Pharm Des* 13(34):3494–3504. <https://doi.org/10.2174/138161207782794257>
13. Gramatica P (2007) *QSAR Comb Sci* 26(5):694–701. <https://doi.org/10.1002/qsar.200610151>
14. Chirico N, Gramatica P (2011) *J Chem Inf Model* 51(9):2320–2335. <https://doi.org/10.1021/ci200211n>
15. Roy PP, Leonard JT, Roy K (2008) *Chemom Intell Lab Syst* 90(1):31–42. <https://doi.org/10.1016/j.chemolab.2007.07.004>
16. Roy K, Mitra I, Ojha PK, Kar S, Das RN, Kabir H (2012) *Chemom Intell Lab Syst* 118:200–210. <https://doi.org/10.1016/j.chemolab.2012.06.004>
17. Masand VH, Mahajan DT, Nazeruddin GM, Hadda TB, Rastija V, Alfeefy AM (2015) *Med Chem Res* 24(3):1241–1264. <https://doi.org/10.1007/s00044-014-1193-8>
18. Ghaemian P, Shayanfar A (2017) *Lett Drug Des Discov* 14(9):999–1007. <https://doi.org/10.2174/1570180814666170126150447>
19. Toropova AP, Toropov AA (2019) *Curr Top Med Chem* 19(29):2643–2657. <https://doi.org/10.2174/1568026619666191105111817>
20. Martin TM, Harten P, Young DM, Muratov EN, Golbraikh A, Zhu H, Tropsha A (2012) *J Chem Inf Model* 52(10):2570–2578. <https://doi.org/10.1021/ci300338w>
21. Toropov AA, Toropova AP, Puzyn T, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2013) *Chemosphere* 92(1):31–37. <https://doi.org/10.1016/j.chemosphere.2013.03.012>
22. Toropova AP, Toropov AA, Benfenati E, Leszczynska D, Leszczynski J (2015) *Bioorg Med Chem* 23(6):1223–1230. <https://doi.org/10.1016/j.bmc.2015.01.055>
23. Toropov AA, Toropova AP (2019) *Struct Chem* 30(5):1677–1683. <https://doi.org/10.1007/s11224-019-01361-6>
24. Hemmateenejad B, Javidnia K, Miri R, Elyasi M (2012) *J Iran Chem Soc* 9(1):53–60. <https://doi.org/10.1007/s13738-011-0005-z>
25. Shayanfar A, Shayanfar S (2014) *Eur J Pharm Sci* 59(1):31–35. <https://doi.org/10.1016/j.ejps.2014.03.007>
26. Consonni V, Ballabio D, Todeschini R (2009) *J Chem Inf Model* 49(7):1669–1678. <https://doi.org/10.1021/ci9000115y>
27. Roy K, Kar S (2014) *Eur J Pharm Sci* 62:111–114. <https://doi.org/10.1016/j.ejps.2014.05.019>
28. Lin LI-K (1992) *Biometrics* 48(2):599–604. <https://doi.org/10.2307/2532314>
29. Toropov AA, Toropova AP (2017) *Mutat Res Genet Toxicol Environ Mutagen* 819:31–37. <https://doi.org/10.1016/j.mrgentox.2017.05.008>
30. Toropov AA, Toropova AP (2020) *Sci Total Environ* 737:139720. <https://doi.org/10.1016/j.scitotenv.2020.139720>
31. Doweiko AM (2008) *IDrugs* 11(12):894–899
32. Skinnider MA, Stacey RG, Wishart DS, Foster LJ (2021) *Nat Mach Intell* 3(9):759–770. <https://doi.org/10.1038/s42256-021-00368-1>
33. Chen J-H, Tseng YJ (2021) *Brief Bioinform* 22(3):bbaa092. <https://doi.org/10.1093/bib/bbaa092>
34. Redman L, Friman M, Gärling T, Hartig T (2013) *Transp Policy* 25:119–127. <https://doi.org/10.1016/j.tranpol.2012.11.005>
35. Zadeh LA (1965) *Inf Control* 8(3):338–353. [https://doi.org/10.1016/S0019-9958\(65\)90241-X](https://doi.org/10.1016/S0019-9958(65)90241-X)
36. Jafari K, Fatemi MH, Toropova AP, Toropov AA (2020) *Chem Phys Lett* 754:137614. <https://doi.org/10.1016/j.cplett.2020.137614>

37. Ahmadi S, Toropova AP, Toropov AA (2020) *Nanotoxicology* 14(8):1118–1126. <https://doi.org/10.1080/17435390.2020.1808252>
38. Toropov AA, Sizochenko N, Toropova AP, Leszczynska D, Leszczynski J (2020) *J Mol Liq* 317:113929. <https://doi.org/10.1016/j.molliq.2020.113929>
39. Toropova AP, Toropov AA (2020) *Fuller Nanotub Carbon Nanostructures* 28(11):900–906. <https://doi.org/10.1080/1536383X.2020.1779705>
40. Kumar P, Kumar A (2021) *J Mol Struct* 1246:131205. <https://doi.org/10.1016/j.molstruc.2021.131205>
41. Jafari K, Fatemi MH, Toropova AP, Toropov AA (2022) *Chemom Intell Lab Syst* 222:104500. <https://doi.org/10.1016/j.chemolab.2022.104500>
42. Kumar P, Kumar A, Singh D (2022) *Environ Toxicol Pharmacol* 93:103893. <https://doi.org/10.1016/j.etap.2022.103893>
43. Toropov AA, Toropova AP (2018) *Anticancer Res* 38(11):6189–6194. <https://doi.org/10.21873/anticancerres.12972>
44. Toropova AP, Toropov AA (2018) *Environ Sci Pollut Res* 25(31):31771–31775. <https://doi.org/10.1007/s11356-018-3291-5>
45. Toropova MA, Raškova M, Raška I Jr, Toropova AP (2019) *Monatsh fur Chem* 150(4):617–623. <https://doi.org/10.1007/s00706-019-2368-2>
46. Ahmadi S (2020) *Chemosphere* 242:125192. <https://doi.org/10.1016/j.chemosphere.2019.125192>
47. Nimbhal M, Bagri K, Kumar P, Kumar A (2020) *Struct Chem* 31(2):831–839. <https://doi.org/10.1007/s11224-019-01468-w>
48. Kumar P, Kumar A (2020) *SAR QSAR Environ Res* 31(9):697–715. <https://doi.org/10.1080/1062936X.2020.1806105>
49. Toropov AA, Toropova AP (2021) *Toxicol Lett* 340:133–140. <https://doi.org/10.1016/j.toxlet.2021.01.015>
50. Toropov AA, Toropova AP, Benfenati E (2010) *Eur J Med Chem* 45(9):3581–3587. <https://doi.org/10.1016/j.ejmech.2010.05.002>
51. Toropov AA, Toropova AP (2021) *Nanotoxicology* 15(7):995–1004. <https://doi.org/10.1080/17435390.2021.1951387>
52. Toropova AP, Toropov AA (2021) *Int J Environ Res* 15(4):709–722. <https://doi.org/10.1007/s41742-021-00346-w>
53. Toropov AA, Toropova AP, Roncaglioni A, Benfenati E (2021) *New J Chem* 45(44):20713–20720. <https://doi.org/10.1039/d1nj03394h>
54. Toropov AA, Toropova AP, Achary PGR, Raškova M, Raška I (2022) *Toxicol Mech Methods* (in press). <https://doi.org/10.1080/15376516.2022.2053918>
55. Toropova AP, Toropov AA, Roncaglioni A, Benfenati E (2022) *Chem Phys Lett* 790:139354. <https://doi.org/10.1016/j.cplett.2022.139354>
56. Toropov AA, Toropova AP, Kudyshkin VO (2022) *Struct Chem* 33(2):617–624. <https://doi.org/10.1007/s11224-021-01875-y>

Chapter 17

CORAL: Predictions of Quality of Rice Based on Retention Index Using a Combination of Correlation Intensity Index and Consensus Modelling



Parvin Kumar and Ashwani Kumar

Abstract The purpose of this study is to utilize the Monte Carlo technique of CORAL software for establishing a quantitative structure-retention relationship (QSRR) for the retention indices of 136 primary flavour volatile organic molecules. SMILES notations of volatile organic compounds were used to compute the descriptor of correlation weight (DCW). Eight splits have been constructed from the dataset of 136 volatile organic chemicals, each of which was further divided into four sets: training, invisible training, calibration and validation. Two target functions i.e. TF₁ (CII_{weight} = 0.0), TF₂ (CII_{weight} = 0.3) were applied to build 16 QSRR models. All QSRR models were statistically good. The coefficient of determination derived by TF₂ for the validation set of split 4 has the maximum statistical result ($R^2_{\text{validation}} = 0.9532$), hence it was accepted as the best model. The assignment of correlation intensity index (CII) on QSPR models was thoroughly examined and found to be more consistent and relevant. The common promoters of increase and decrease of endpoint were also extracted from four splits 1, 2, 3 and 4. Furthermore, consensus modelling using the split 4 architecture of dataset distribution enhances prediction accuracy by increasing the numerical value of $R^2_{\text{validation}}$ from 0.9532 to 0.9864.

Keywords QSPR/QSAR · Retention index · Validation · Correlation intensity index

P. Kumar (✉)

Department of Chemistry, Kurukshetra University, Kurukshetra, Haryana 136119, India
e-mail: parvinchem@kuk.ac.in

A. Kumar

Department of Pharmaceutical Sciences, Guru Jambheshwar University of Science and Technology, Hisar, Haryana 125001, India

17.1 Introduction

Rice (*Oryza sativa* L.) is not only the primary source of calories for almost half of the world's population but also offers food security to many low-income nations. In many Asian countries, rice is the most important staple cereal after wheat and maize, accounting for more than 90% of rice consumption worldwide [1]. White rice is the most popular type of rice, and these grains are progressively ground from the outside to the interior, resulting in disparities in composition between the rice layers, which affects functional and edible qualities [2]. From the outer surface to the inner side of the rice grains, the content of protein, fat, mineral and other non-starch components dropped, while the content of amylose and starch enhances [3, 4]. As a result, crop quality management must be improved to confirm that rice has the best organoleptic features and is admissible to consumers.

The colour, texture, unique taste and aromatic composition of different rice varieties can be used to identify the grain quality of the rice. Moreover, it has been noticed that modest alterations in sensory qualities, particularly aroma, alter customer acceptability. Currently, more than 150 diverse volatile substances have been identified in rice, primarily from the alkanes, alcohols, phenols, aldehydes, ketones, enones, furanone, fatty acids, esters, benzyl derivatives, monoterpenoids, sesquiterpenoids, naphthalenes, xylenes, furans, pyridines and pyrroles [5–7]. Generally, gas chromatography (GC) is applied to analyse the volatile compounds present in rice.

The quantitative structure–property relationship (QSPR) is a method that uses “*descriptor-property*” correlations to estimate unknown numerical data on endpoints of significance. After the innovative investigations of the use of QSPR to chromatographic retention indices (I), investigators are more interested in using the quantitative structure retention relationships method (QSRR) [8]. The models generated by the QSRR method can be applied to predict the retention index of unknown compounds and to separate the complex chemical mixtures [8, 9]. In the recent decade, The CORrelation And Logic (CORAL) programme has been recommended as a useful tool in QSAR/QSPR experiments (<http://www.insilico.eu/CORAL>) [10–12]. A global molecular descriptor, i.e. a descriptor of correlation weight (DCW) calculated by CORAL software, is applied to anticipate or compute the required endpoint value [13–15]. The inbuilt Monte Carlo algorithm of CORAL software is used to compute the correlation between the DCW and endpoints. Recently, the correlation intensity index (CII) is also applied to generate better QSPR models [16–24]. The goal of QSAR researchers is to improve prediction accuracy by attaining minimal predicted residuals for test molecules. When diverse models are present, a few QSPR scientists have adopted a consensus method of modelling to attain this purpose [25–27]. Various researchers have employed the original consensus modelling by considering all the separately developed models having good statistical results [28]. Consensus modelling is seen to be superior to individual models because it contains all of the information found in many individual models. Roy et al. have created an intriguing “*Intelligent Consensus Prediction*” programme for performing consensus modelling [29].

Because of the aforementioned information and as part of our ongoing efforts to develop QSPR models, the goal of the present research is to construct robust QSRR models for the retention indices of 136 volatile organic compounds (VOC) detected in the headspace of rice. Here, we have also implemented the correlation intensity index (CII), a new predictive potential criterion, and consensus modelling to get a better predictive model.

17.2 Materials and Method

17.2.1 Data

Experimental retention indices (RI) for 136 major flavour volatile molecules were collected from the literature [9]. The Divinylbenzene-Carboxen-Polydimethylsiloxane (DVB-CAR-PDMS) fibre was used to determine the experimental retention indices using solid-phase microextraction-gas chromatography-mass spectrometry (SPME-GC-MS). ChemAxon was employed to draw chemical structures, which were then converted into SMILES notation [30]. All SMILES notations were converted into canonical SMILES using the Open Babel programme [31]. The molecule 2,2,4-trimethylheptane (Cas Number 14720-74-2) was found as a duplication of trimethylheptane during the database screening. As a result, the trimethylheptane molecule was eliminated, and the retention index for 2,2,4-trimethylheptane was taken as 878.5 (average of two RI). The molecular formula of the compound 2,6-bis-(*t*-butyl)-2,5-cyclohexadien-1-one (CAS number 6378-27-8) did not match with its structural formula, therefore it was not considered for QSRR model development. As a result, the QSRR model was built using 136 chemicals. The dataset of 136 major flavour volatile molecules is categorized into four sets: active training (for building the model), passive training (for inspecting the constructed model on the entities not included in the active training set), calibration (to monitor overtraining) and validation (for validating the model's external predictability) (Table 17.1A, B). The whole dataset was used to make eight random splits and these were non-identical in line with the mathematical equation given in the literature [32].

17.2.2 Model

The following equation is used for the prediction of RI of VOCs:

$$RI = C_0 + C_1 \times^{\text{SMILES}} \text{DCW}(T^*, N_{\text{epoch}}^*) \quad (17.1)$$

Table 17.1 A The CAS number, SMILES notation and split distribution

S. No.	CAS number	Split								SMILES	
		1	2	3	4	5	6	7	8		
1	64-17-5	AT	PT	CL	VD	CL	VD	VD	AT	PT	CCO
2	67-64-1	PT	AT	VD	CL	AT	PT	PT	CL	VD	CC(=O)C
3	75-18-3	CL	VD	AT	PT	CL	VD	VD	AT	PT	CSC
4	110-54-3	VD	CL	PT	AT	AT	PT	PT	CL	VD	CCCCC
5	123-72-8	AT	PT	CL	VD	VD	CL	CL	PT	AT	CCCC=O
6	64-19-7	PT	AT	VD	CL	PT	AT	AT	VD	CL	CC(=O)O
7	590-86-3	VD	CL	PT	AT	PT	AT	AT	VD	CL	O=CCC(C)C
8	96-17-3	AT	PT	CL	VD	VD	VD	PT	CL	VD	CC(C=O)CC
9	142-82-5	PT	AT	VD	CL	CL	VD	VD	AT	PT	CCCCCCC
10	110-62-3	CL	VD	AT	PT	AT	CL	CL	PT	AT	CCCCC=O
11	623-42-7	VD	CL	PT	AT	CL	VD	VD	AT	PT	CCCC(=O)OC
12	123-51-3	AT	PT	CL	VD	VD	PT	AT	VD	CL	OCCC(C)C
13	137-32-6	PT	AT	VD	CL	VD	CL	VD	PT	AT	CCC(CO)C
14	624-92-0	CL	VD	AT	PT	PT	PT	PT	CL	VD	CSSC
15	110-86-1	PT	AT	VD	CL	AT	PT	PT	CL	VD	c1cccnc1
16	589-38-8	VD	CL	PT	AT	VD	VD	CL	PT	AT	CCCC(=O)CC
17	71-41-0	AT	PT	CL	VD	CL	VD	VD	AT	PT	CCCCCO
18	108-88-3	CL	VD	AT	PT	CL	AT	AT	VD	CL	Cc1ccccc1
19	123-54-6	VD	CL	PT	AT	AT	PT	PT	CL	VD	CC(=O)CC(=O)C
20	513-85-9	AT	PT	CL	VD	VD	VD	CL	PT	AT	CC(C(O)C)O

(continued)

Table 17.1 A (continued)

S. No.	CAS number	Split								SMILES
		1	2	3	4	5	6	7	8	
21	107-88-0	PT	AT	VD	CL	PT	AT	VD	CL	OCCCC(O)C
22	111-65-9	VD	CL	PT	AT	PT	AT	VD	CL	CCCCCCCC
23	66-25-1	CL	VD	AT	PT	VD	VD	AT	PT	CCCCC=O
24	624-24-8	AT	PT	CL	VD	AT	PT	CL	VD	CCCCC(=O)OC
25	100-40-3	PT	AT	VD	CL	CL	VD	AT	PT	C=CC1CCCCC1
26	107-92-6	CL	VD	AT	PT	VD	CL	PT	AT	CCCC(=O)O
27	6728-26-3	CL	VD	AT	PT	AT	PT	CL	VD	CCC/C=C/C=O
28	6789-80-6	VD	CL	PT	AT	CL	VD	AT	PT	CC/C=C\C/C=O
29	100-41-4	AT	PT	CL	VD	PT	AT	VD	CL	Cc1cccc1
30	111-27-3	PT	AT	VD	CL	VD	CL	PT	AT	CCCCCO
31	106-42-3	CL	VD	AT	PT	PT	AT	VD	CL	Cc1ccc(cc1)C
32	109-52-4	VD	CL	PT	AT	VD	CL	PT	AT	CCCCC(=O)O
33	14720-74-2	AT	PT	CL	VD	CL	VD	AT	PT	CCCC(CC(C)C)C)C
34	110-43-0	PT	AT	VD	CL	AT	PT	CL	VD	CCCCC(=O)C
35	4466-24-4	CL	VD	AT	PT	CL	VD	AT	PT	CCCCc1cccc1
36	100-42-5	VD	CL	PT	AT	AT	PT	CL	VD	C=Cc1cccc1
37	94-47-6	AT	PT	CL	VD	VD	CL	PT	AT	Cc1cccc1C
38	111-84-2	PT	AT	VD	CL	PT	AT	VD	CL	CCCCCCCC
39	111-71-7	CL	VD	AT	PT	VD	CL	PT	AT	CCCCC=O

(continued)

Table 17.1 A (continued)

S. No.	CAS number	Split								SMILES
		1	2	3	4	5	6	7	8	
40	111-76-2	VD	CL	PT	AT	PT	AT	VD	CL	CCCCOCCO
41	4032-93-3	AT	PT	CL	VD	AT	VD	CL	VD	CC(CCC(C(C)C)C)C
42	85213-22-05	PT	AT	VD	CL	CL	VD	AT	PT	CC(=O)C1=NCCCC1
43	106-70-7	CL	VD	AT	PT	AT	PT	CL	VD	CCCCC(=O)OC
44	80-56-8	VD	CL	PT	AT	CL	VD	AT	PT	CC1=CCC2CC1C2(C)C
45	5131-66-8	AT	PT	CL	VD	PT	AT	VD	CL	CCCCOCC(O)C
46	57266-86-1	PT	AT	VD	CL	VD	CL	PT	AT	CCCC/C=C/C=O
47	611-14-3	CL	VD	AT	PT	PT	AT	VD	CL	CCc1ccc(cc1)C
48	100-52-7	VD	CL	PT	AT	VD	CL	PT	AT	O=Cc1ccccc1
49	111-70-6	AT	PT	CL	VD	CL	VD	AT	PT	CCCCCCCCO
50	3658-80-8	PT	AT	VD	CL	AT	PT	CL	VD	CSSC
51	142-62-1	VD	CL	PT	AT	AT	PT	CL	VD	CCCCC(=O)O
52	110-93-0	CL	VD	AT	PT	CL	VD	AT	PT	CC(=O)CC/C=C(C)C
53	3777-69-3	PT	AT	VD	CL	PT	AT	VD	CL	CCCCCclccc1
54	1462-84-6	AT	PT	CL	VD	VD	CL	PT	AT	Cc1ccc(c(n1)C)C
55	95-63-6	CL	VD	AT	PT	VD	CL	PT	AT	Cc1ccc(c(c1)C)C
56	123-66-0	VD	CL	PT	AT	PT	AT	VD	CL	CCCCC(=O)OCC
57	124-18-5	AT	PT	CL	VD	AT	PT	CL	VD	CCCCCCCCC
58	124-13-0	PT	AT	VD	CL	CL	VD	AT	PT	CCCCCCCC=O

(continued)

Table 17.1 A (continued)

S. No.	CAS number	Split								SMILES
		1	2	3	4	5	6	7	8	
59	104-76-7	CL	VD	AT	PT	AT	PT	CL	VD	CCCC(CO)CC
60	138-86-3	VD	CL	PT	AT	CL	VD	AT	PT	CC1=CCC(CC1)C(=O)C
61	496-11-7	AT	PT	CL	VD	PT	AT	VD	CL	C1Cc2c(Cl)cccc2
62	100-51-6	PT	AT	VD	CL	VD	CL	PT	AT	OCc1cccc1
63	18402-82-9	CL	VD	AT	PT	PT	AT	VD	CL	CCCC/C=C/C(=O)C
64	122-78-1	VD	CL	PT	AT	VD	CL	PT	AT	O=CCc1cccc1
65	695-06-7	AT	PT	CL	VD	CL	VD	AT	PT	CCC1CCC(=O)O1
66	2548-87-0	PT	AT	VD	CL	AT	PT	CL	VD	CCCCC/C=C/C=O
67	111-87-5	CL	VD	AT	PT	CL	VD	AT	PT	CCCCCCCCO
68	768-49-0	VD	CL	PT	AT	AT	PT	CL	VD	C/C(=C\c1cccc1)/C
69	527-84-4	AT	PT	CL	VD	VD	CL	PT	AT	CC(c1cccc1C)C
70	1120-21-4	PT	AT	VD	CL	PT	AT	VD	CL	CCCCCCCCCCC
71	124-19-6	CL	VD	AT	PT	VD	CL	PT	AT	CCCCCCCCC=O
72	488-23-3	VD	CL	PT	AT	PT	AT	VD	CL	Cc1c(C)ccc(c1C)C
73	111-11-5	AT	PT	CL	VD	AT	PT	CL	VD	CCCCCCCC(=O)OC
74	57283-79-1	PT	AT	VD	CL	CL	VD	AT	PT	CCC(C(C)C)/C=C/C(=O)C
75	72218-58-7	CL	VD	AT	PT	AT	PT	CL	VD	CCCC(C(OC(=O)C)C)C
76	105-21-5	VD	CL	PT	AT	CL	VD	AT	PT	CCCC1CCC(=O)O1
77	138-87-4	AT	PT	CL	VD	PT	AT	VD	CL	CC(=O)C1CCCC(CC1)O

(continued)

Table 17.1 A (continued)

S. No.	CAS number	Split								SMILES		
		1	2	3	4	5	6	7	8			
78	18829-56-6	PT	AT	VD	CL	VD	CL	VD	CL	PT	AT	CCCCC/C=C/C=O
79	143-08-8	CL	VD	AT	PT	PT	AT	VD	PT	VD	CL	CCCCCCCCCO
80	91-20-3	VD	CL	PT	AT	VD	AT	VD	CL	PT	AT	c1ccc2c(c1)ccc2
81	106-32-1	AT	PT	CL	VD	CL	VD	CL	VD	AT	PT	CCCCCCCC(=O)OCC
82	98-55-5	PT	AT	VD	CL	AT	PT	AT	PT	CL	VD	CC1=CCC(CC1)C(O)(C)C
83	112-40-3	CL	VD	AT	PT	AT	PT	CL	VD	AT	PT	CCCCCCCCCCCC
84	586-81-2	VD	CL	PT	AT	AT	AT	AT	PT	CL	VD	C/C(=C)/CCC(CC1)(C)O)/C
85	112-31-2	AT	PT	CL	VD	CL	VD	VD	CL	PT	AT	CCCCCCCCCCC=O
86	5910-87-2	PT	AT	VD	CL	VD	CL	PT	AT	VD	CL	CCCC/C=C/C=C/C=O
87	85-16-9	CL	VD	AT	PT	AT	PT	VD	CL	PT	AT	c1ccc2c(c1)scn2
88	19780-79-1	VD	CL	PT	AT	PT	AT	PT	AT	VD	CL	CCCCCCC(CCCCC)CO
89	104-50-7	AT	PT	CL	VD	CL	VD	AT	PT	CL	VD	CCCCC1CCC(=O)O1
90	3913-81-3	PT	AT	VD	CL	VD	CL	CL	VD	AT	PT	CCCCCCC/C=C/C=O
91	2437-56-1	CL	VD	AT	PT	AT	PT	AT	PT	CL	VD	CCCCCCCCCCCC=C
92	585-34-2	VD	CL	PT	AT	CL	AT	CL	VD	AT	PT	Oc1ccc(c1)C(C)(C)C
93	123-29-5	AT	PT	CL	VD	CL	VD	PT	AT	VD	CL	CCCCCCCCC(=O)OCC
94	112-12-9	CL	VD	AT	PT	AT	PT	PT	AT	VD	CL	CCCCCCCCCCCC(=O)C
95	120-72-9	PT	AT	VD	CL	VD	CL	VD	CL	PT	AT	c1ccc2c(c1)fnHlcc2
96	629-50-5	VD	CL	PT	AT	VD	AT	VD	CL	PT	AT	CCCCCCCCCCCCCCC

(continued)

Table 17.1 A (continued)

S. No.	CAS number	Split								SMILES	
		1	2	3	4	5	6	7	8		
97	112-44-7	AT	PT	CL	VD	CL	VD	AT	VD	PT	CCCCCCCCCCCC=O
98	110-42-9	PT	AT	VD	CL	AT	CL	PT	AT	VD	CCCCCCCCCCC(=O)OC
99	104-61-0	CL	VD	AT	PT	CL	PT	AT	VD	PT	CCCCC1CCC(=O)O1
100	13019-16-4	VD	CL	PT	AT	AT	PT	CL	VD	VD	CCCCC/C=C/(CCCC)C=O
101	1120-36-1	AT	PT	CL	VD	VD	VD	CL	PT	AT	CCCCCCCCCCCCC=C
102	110-38-3	PT	AT	VD	CL	PT	AT	VD	CL	CL	CCCCCCCCCCC(=O)OCC
103	629-59-4	CL	VD	AT	PT	VD	VD	CL	PT	AT	CCCCCCCCCCCCC
104	1135-66-6	VD	CL	PT	AT	PT	AT	VD	AT	CL	CC1(C)C2CCC3(C1=CCCC3(C)C)C2
105	87-44-5	AT	PT	CL	VD	CL	VD	CL	VD	VD	C/C1=C/C/C(=C)[C@@H]2[C@@H](CC1)C(C2)(C)C
106	6938-94-9	PT	AT	VD	CL	CL	CL	VD	AT	PT	CC(OC(=O)CCCC(=O)OC(C)C)C
107	3879-26-3	CL	VD	AT	PT	AT	PT	CL	VD	VD	C/C(=C)CCC(=O)C)CC/C=C(O)C
108	719-22-2	VD	CL	PT	AT	CL	VD	AT	VD	PT	O=C1C=C(C(=O)C(=C)1)C(C)C(C)C(C)C
109	629-73-2	AT	PT	CL	VD	PT	VD	AT	VD	CL	CCCCCCCCCCCCCCC=C
110	629-62-9	PT	AT	VD	CL	VD	CL	PT	CL	AT	CCCCCCCCCCCCCCC
111	128-37-0	CL	VD	AT	PT	PT	AT	VD	AT	CL	Cc1cc(c(c1)C(C)C)O(C)C(C)C
112	111-82-0	VD	CL	PT	AT	VD	VD	CL	PT	AT	CCCCCCCCCCCCC(=O)OC
113	483-77-2	AT	PT	CL	VD	CL	VD	AT	VD	PT	CC([C@@H]1CC[C@@H](c2c1cc(C)cc2)C)C
114	4130-42-1	PT	AT	VD	CL	AT	CL	PT	CL	VD	CCc1cc(c(c1)C(C)C)O(C)C(C)C
115	106-33-2	CL	VD	AT	PT	CL	VD	AT	VD	PT	CCCCCCCCCCCCC(=O)OCC

(continued)

Table 17.1 A (continued)

S. No.	CAS number	Split								SMILES
		1	2	3	4	5	6	7	8	
116	544-76-3	VD	CL	PT	AT	AT	PT	CL	VD	CCCCCCCCCCCCCCCC
117	3892-00-0	AT	PT	CL	VD	VD	CL	PT	AT	CCCCC(CCCCC(CCC(C)O)C)
118	629-78-7	PT	AT	VD	CL	CL	PT	AT	VD	CCCCCCCCCCCCCCCC
119	1921-70-6	CL	VD	AT	PT	VD	VD	CL	PT	CC(CCCC(C)CCCC(CCCC(C)C)C
120	124-10-7	VD	CL	PT	AT	PT	AT	VD	CL	CCCCCCCCCCCC(=O)OC
121	3910-35-8	AT	PT	CL	VD	AT	PT	CL	VD	CC1(CC(e2e1eccc2)(C)C)e1eccc1
122	24157-81-1	PT	AT	VD	CL	CL	VD	AT	PT	CC(e1eccc2e1)ccc(e2)C(C)C
123	31516-55-9	CL	VD	AT	PT	AT	PT	CL	VD	CC(e1eccc1)CC(e1eccc1)(C)C
124	124-06-1	VD	CL	PT	AT	CL	VD	AT	PT	CCCCCCCCCCCC(=O)OCC
125	593-45-3	AT	PT	CL	VD	PT	AT	VD	CL	CCCCCCCCCCCCCCCC
126	638-36-8	PT	AT	VD	CL	VD	CL	PT	AT	CCC(CCCC(CCCC(C)C)C)C
127	502-69-2	CL	VD	AT	PT	PT	AT	VD	CL	CC(CCCC(C)CCCC(CCCC(=O)C)C
128	629-92-5	VD	CL	PT	AT	VD	CL	PT	AT	CCCCCCCCCCCCCCCC
129	112-39-0	AT	PT	CL	VD	CL	VD	AT	PT	CCCCCCCCCCCC(=O)OC
130	628-97-7	PT	AT	VD	CL	AT	PT	CL	VD	CCCCCCCCCCCC(=O)OCC
131	57-10-3	CL	VD	AT	PT	CL	VD	AT	PT	CCCCCCCCCCCC(=O)O
132	112-95-8	VD	CL	PT	AT	AT	PT	CL	VD	CCCCCCCCCCCCCCCC
133	2566-97-4	AT	PT	CL	VD	VD	CL	PT	AT	CCCCC=C/C/C=C/C/CCCC(=O)OC
134	112-62-9	PT	AT	VD	CL	PT	AT	VD	CL	CCCCCCCC=C\CCCCCCCC(=O)OC
135	544-35-4	CL	VD	AT	PT	VD	CL	PT	AT	CCCCC=C\C/C/C=C\CCCCCCCC(=O)OCC

(continued)

Table 17.1 A (continued)

S. No.	CAS number	Split								SMILES
		1	2	3	4	5	6	7	8	
136	111-62-6	VD	CL	PT	AT	PT	AT	VD	CL	CCCCCCCC/C=C\CCCCCCCC(=O)OCC

AT is the active training set; PT is a passive training set; CL is the calibration set; VD is the validation set

Table 17.1 B The experimental and calculated retention indices (RI) using TF₂

S. No.	Expr RI	Calculated RI							
		Split 1	Split 2	Split 3	Split 4	Split 5	Split 6	Split 7	Split 8
137	250	423.3632	437.1495	389.5553	483.8953	426.5486	409.7672	387.9183	466.8564
138	259	385.4168	390.8645	477.4659	471.4841	448.8875	477.9003	297.4013	510.5662
139	308	528.5303	557.61	307.6108	654.4359	452.8354	849.0861	309.6736	393.9715
140	600	652.3318	667.4791	603.9851	645.2152	614.5331	647.8272	600.1631	674.4936
141	602	622.8083	636.9365	581.4894	640.4943	587.7205	602.2795	561.2641	646.0716
142	622	508.8817	581.6629	472.8613	545.9045	483.0414	511.5518	467.5713	540.2986
143	652	713.8693	684.303	623.2411	653.2522	605.2172	694.8382	559.7057	593.2393
144	660	643.0949	642.3981	591.1918	633.5451	637.7606	597.9549	533.8783	595.4547
145	700	746.9076	760.3405	710.7586	739.078	714.4143	741.3659	707.6098	769.227
146	701	717.3841	729.7979	688.263	734.3571	687.6017	695.8183	668.7108	740.805
147	710	758.5273	750.3526	716.442	733.9296	742.423	737.133	697.7993	758.5131
148	730	727.1113	764.1248	865.9935	684.3787	651.0646	759.4184	571.8832	744.6657
149	730	725.8811	641.1598	673.4496	644.5871	599.7454	628.0169	669.8098	723.4327
150	741	528.5303	741.3812	738.063	654.4359	688.0205	849.0861	708.1232	393.9715
151	763	659.656	762.8145	883.025	809.9762	765.7621	695.2896	558.4108	724.166
152	744	652.0235	750.7673	733.3521	782.5355	788.177	757.5809	636.2201	783.3345
153	761	707.0907	715.7337	709.876	765.4836	726.1921	690.3835	710.2585	751.0567
154	764	774.1594	887.5483	802.3636	829.8071	741.1552	843.387	839.146	871.4956
155	778	760.8541	835.2007	725.0704	778.9722	778.9819	769.3507	614.699	818.09
156	788	774.1207	849.5683	722.9876	857.7766	819.4118	705.1657	741.3498	761.892

(continued)

Table 17.1 B (continued)

S. No.	Expr RI	Calculated RI							
		Split 1	Split 2	Split 3	Split 4	Split 5	Split 6	Split 7	Split 8
157	791	778.0926	762.8952	668.5601	776.0945	719.4916	765.7998	673.6627	751.8022
158	800	841.4834	853.2019	817.5322	832.9408	814.2954	834.9047	815.0565	863.9604
159	800	811.96	822.6593	795.0365	828.2198	787.4828	789.3571	776.1576	835.5384
160	821	853.1031	843.214	823.2156	827.7924	842.3042	830.6717	805.246	853.2466
161	832	811.2288	886.4857	916.978	906.9867	802.2533	925.1571	843.8723	924.2686
162	638	698.0334	767.3856	686.4085	733.63	682.8037	698.6294	682.4648	729.7655
163	850	851.7479	879.2148	870.9236	890.1945	875.6179	865.4935	850.3559	894.6059
164	853	775.5178	830.7402	795.5264	768.289	894.4705	840.5825	836.4372	821.1603
165	858	817.9016	864.9905	897.8801	878.7657	838.3559	858.3113	874.7317	936.5689
166	865	801.6666	808.5951	816.6495	859.3464	826.0733	783.9222	817.7053	845.7901
167	868	897.7982	1097.662	873.7345	902.7712	859.1824	939.5412	906.0407	937.7687
168	879	792.6092	860.247	793.182	827.4928	782.6849	792.1681	789.9115	824.4989
169	883	836.7828	861.2951	765.1554	694.6839	920.9488	804.1093	848.3963	924.7495
170	888	763.7201	762.3101	904.5601	846.9352	848.4122	852.0553	727.1882	889.4999
171	890	809.7551	895.3403	891.8237	903.7914	834.0881	892.7772	888.2782	973.2048
172	892	785.9593	921.8187	874.2337	859.7552	893.793	846.1933	881.8735	915.7924
173	893	880.8299	868.1528	841.3168	789.5186	782.2703	801.5603	875.2425	871.103
174	900	936.0593	946.0633	924.3057	926.8035	914.1766	928.4434	922.5033	958.6938
175	902	906.5358	915.5207	901.8101	922.0826	887.364	882.8958	883.6043	930.2718

(continued)

Table 17.1 B (continued)

S. No.	Expr RI	Calculated RI							
		Split 1	Split 2	Split 3	Split 4	Split 5	Split 6	Split 7	Split 8
176	905	771.7452	822.8348	829.598	875.7375	840.3154	835.8494	665.5337	800.6541
177	913	936.9595	907.2748	928.8326	822.9964	836.5379	890.9762	823.1698	921.415
178	920	840.9174	920.8337	741.9567	949.9821	903.3432	723.5854	920.3138	954.263
179	922	947.679	936.0754	929.9891	921.6552	942.1854	924.2105	912.6928	947.98
180	932	707.6388	982.4978	944.1149	924.4422	926.9763	1115.123	936.9886	991.4441
181	938	942.9206	942.2178	910.0424	958.2421	942.5659	993.264	809.5072	956.3679
182	956	946.3238	972.0762	977.6972	984.0572	975.4991	959.0323	957.8027	989.3393
183	959	941.5404	1075.104	969.2509	951.7298	956.3831	954.4655	941.6264	1002.842
184	962	899.955	1066.353	964.6061	980.8703	1025.813	978.6551	936.4088	916.2785
185	969	896.2424	901.4565	923.4231	953.2091	925.9544	877.461	925.152	940.5236
186	970	528.5303	968.0826	869.6219	654.4359	968.8739	849.0861	907.2824	393.9715
187	983	887.1851	953.1084	899.9556	921.3556	882.566	885.7069	897.3582	919.2323
188	983	884.5502	924.1293	994.5922	1001.828	1006.334	971.9209	979.028	977.9127
189	990	904.331	988.2017	998.5973	997.6542	933.9692	986.3159	995.7249	1067.938
190	989	896.9086	991.1329	1014.139	816.6905	865.159	766.6403	893.2677	904.5882
191	994	1013.347	965.8065	930.4816	1018.116	837.9637	998.9617	1011.367	989.7757
192	998	1024.639	1068.588	1008.718	1020.272	1026.382	1029.513	950.2736	1033.368
193	1000	1030.635	1038.925	1031.079	1020.666	1014.058	1021.982	1029.95	1053.427
194	1004	1001.112	1008.382	1008.584	1015.945	987.2452	976.4346	991.051	1025.005

(continued)

Table 17.1 B (continued)

S. No.	Expr RI	Calculated RI							
		Split 1	Split 2	Split 3	Split 4	Split 5	Split 6	Split 7	Split 8
195	1029	992.4878	1001.063	929.3358	955.6385	939.035	907.6975	1008.629	996.2011
196	1030	1231.57	1028.367	1057.812	1044.374	1012.442	959.4131	1036.029	1015.863
197	1035	1073.994	1049.286	1005.745	1077.19	1002.022	1001.651	1089.299	1042.76
198	1036	960.6403	1034.804	1136.751	1039.601	943.7283	1035.29	1047.52	1083.114
199	1038	950.1482	1037.538	1033.885	986.0698	999.8219	1011.274	923.6228	1095.812
200	1045	999.1333	1079.456	1055.535	1077.446	1006.11	1046.798	1013.971	996.761
201	1049	1064.947	1040.33	1055.532	1030.55	1049.222	1094.705	1045.259	1054.89
202	1058	1040.9	1064.938	1084.471	1077.92	1075.38	1052.571	1065.249	1084.073
203	1071	990.8183	994.3179	1030.197	1047.072	1025.836	970.9998	1032.599	1035.257
204	1082	1086.519	1098.327	1103.341	1067.724	1080.239	1019.761	1107.353	1089.561
205	1084	1075.388	1062.423	1007.892	972.3808	904.4949	1032.915	1074.651	1058.464
206	1100	1125.211	1131.786	1137.853	1114.529	1113.939	1115.521	1137.397	1148.161
207	1106	1095.688	1101.244	1115.357	1109.808	1087.126	1069.973	1098.498	1119.739
208	1123	1121.937	1147.795	1149.492	1089.886	1114.389	1103.917	957.5147	1093.481
209	1128	1136.831	1121.798	1143.536	1109.381	1141.948	1111.288	1127.586	1137.447
210	1147	1107.481	1152.67	1299.089	1127.813	1157.081	1114.296	1145.602	1154.072
211	1151	1075.751	1128.12	1179.021	1137.42	1113.729	1134.688	1072.392	1185.107
212	1156	1159.523	1133.191	1162.305	1124.413	1149.103	1188.244	1152.706	1149.623
213	1157	1131.036	1165.786	1175.787	1249.616	1152.269	1149.66	1223.48	1191.415

(continued)

Table 17.1 B (continued)

S. No.	Expr RI	Calculated RI							
		Split 1	Split 2	Split 3	Split 4	Split 5	Split 6	Split 7	Split 8
214	1166	1135.476	1157.799	1191.244	1171.783	1175.261	1146.11	1172.696	1178.806
215	1176	1085.394	1087.179	1136.97	1140.935	1125.717	1064.539	1140.046	1129.99
216	1190	1416.42	1190.477	1222.444	1297.608	1144.639	1242.058	1207.156	1274.155
217	1196	1213.791	1254.311	1222.265	1207.998	1226.145	1216.591	1165.167	1222.834
218	1198	1080.744	1195.001	974.063	1092.292	1194.727	1093.416	1181.087	1244.545
219	1200	1219.787	1224.647	1244.626	1208.392	1213.82	1209.06	1244.844	1242.894
220	1203	1232.491	1210.781	1208.718	1224.282	1203.693	1233.516	1199.921	1349.254
221	1207	1190.263	1194.105	1222.131	1203.671	1187.008	1163.512	1205.945	1214.472
222	1218	1148.967	1215.158	1119.275	1135.916	1215.195	1206.883	1220.608	1117.839
223	1234	1129.045	1101.369	1236.341	1144.74	968.4915	1165.723	1198.551	1135.162
224	1254	1527.035	1466.54	1393.596	1506.158	1469.144	1420.732	1751.445	1550.935
225	1259	1254.099	1226.052	1269.079	1218.275	1248.984	1281.783	1260.152	1244.357
226	1267	1230.051	1250.66	1298.018	1265.646	1275.143	1239.649	1280.143	1273.54
227	1293	1261.998	1211.656	1252.115	1239.509	1254.301	1252.807	1296.286	1277.619
228	1295	1186.979	1357.765	1319.605	1245.133	1145.475	1131.03	1291.698	1312.348
229	1296	1308.366	1347.172	1329.039	1301.86	1326.026	1310.13	1272.614	1317.568
230	1300	1142.024	1133.756	1331.654	1222.386	1247.937	1226.21	1156.975	1268.434
231	1300	1314.673	1303.383	1005.075	1184.92	906.9326	1186.105	1319.734	1250.103
232	1300	1314.363	1317.509	1351.4	1302.255	1313.701	1302.599	1352.29	1337.628

(continued)

Table 17.1 B (continued)

S. No.	Expr RI	Calculated RI							
		Split 1	Split 2	Split 3	Split 4	Split 5	Split 6	Split 7	Split 8
233	1310	1284.839	1286.966	1328.904	1297.534	1286.889	1257.051	1313.391	1309.206
234	1325	1325.982	1307.521	1357.083	1297.106	1341.71	1298.366	1342.48	1326.914
235	1368	1348.674	1318.914	1375.852	1312.138	1348.865	1375.322	1367.599	1339.09
236	1375	1327.518	1428.48	1405.675	1354.43	1371.429	1314.397	1484.726	1393.991
237	1392	1356.574	1304.518	1358.889	1333.372	1354.182	1346.346	1403.733	1372.353
238	1394	1402.942	1440.033	1435.813	1395.723	1425.907	1403.668	1380.061	1412.301
239	1400	1408.939	1410.37	1458.174	1396.117	1413.582	1396.137	1459.737	1432.361
240	1408	1332.239	1468.969	1424.824	1423.308	1428.291	1486.96	1222.003	1409.159
241	1433	1438.767	1404.303	1360.546	1438.341	1425.149	1401.382	1568.763	1445.913
242	1448	1394.657	1449.017	1264.19	1378.004	1319.645	1165.702	1445.728	1404.525
243	1450	1384.299	1594.121	1447.597	1379.043	1454.511	1402.805	1590.711	1366.356
244	1471	1644.291	1527.803	1507.636	1460.099	1482.953	1287.017	1475.577	1516.402
245	1492	1545.726	1490.24	1572.436	1521.097	1553.944	1533.423	1618.626	1561.82
246	1500	1503.514	1503.232	1564.947	1489.98	1513.464	1489.676	1567.184	1527.094
247	1510	1527.603	1580.101	1538.28	1451.901	1495.27	1508.34	1711.94	1664.258
248	1523	1515.134	1493.244	1570.631	1484.832	1541.472	1485.443	1557.373	1516.381
249	1534	1508.775	1509.137	1453.284	1272.98	1454.159	1406.125	1535.471	1511.521
250	1561	1571.345	1557.543	1633.796	1500.86	1592.47	1523.265	1747.526	1729.332
251	1592	1592.094	1625.756	1649.36	1583.449	1625.669	1590.746	1594.954	1601.768

(continued)

Table 17.1 B (continued)

S. No.	Expr RI	Calculated RI							
		Split 1	Split 2	Split 3	Split 4	Split 5	Split 6	Split 7	Split 8
252	1600	1598.09	1596.093	1671.721	1583.843	1613.345	1583.215	1674.63	1621.828
253	1646	1677.646	1630.113	1661.522	1540.153	1660.833	1624.609	1705.731	1648.292
254	1700	1692.666	1688.954	1778.494	1677.706	1713.226	1676.754	1782.077	1716.561
255	1703	1729.982	1699.513	1672.235	1566.698	1723.172	1669.275	1757.407	1684.881
256	1724	1704.286	1678.967	1784.178	1672.557	1741.235	1672.521	1772.267	1705.847
257	1731	1746.633	1693.188	1842.38	1666.535	1733.108	1698.891	1983.994	1968.58
258	1744	1768.071	1736.977	1841.475	1576.763	1786.453	1619.403	1731.466	1724.203
259	1762	1613.352	1714.456	1775.072	1668.28	1771.303	1732.788	2033.318	1783.264
260	1793	1781.246	1811.479	1862.907	1771.174	1825.432	1777.823	1809.848	1791.235
261	1800	1787.242	1781.816	1885.268	1771.568	1813.107	1770.292	1889.524	1811.295
262	1807	1824.558	1792.374	1779.009	1660.561	1823.053	1762.814	1864.853	1779.614
263	1844	1677.337	1713.401	1790.889	1677.473	1834.477	1734.363	1741.788	1757.133
264	1900	1881.818	1874.677	1992.041	1865.431	1912.988	1863.831	1996.971	1906.028
265	1925	1893.437	1864.689	1997.725	1860.283	1940.997	1859.598	1987.16	1895.314
266	1993	1970.397	1997.202	2076.454	1958.9	2025.194	1964.901	2024.741	1980.702
267	1995	1832.944	1881.722	1967.691	1859.983	1881.378	1821.095	1971.826	1866.567
268	2000	1976.394	1967.539	2098.815	1959.294	2012.869	1957.37	2104.417	2000.762
269	2051	2101.776	1988.696	1936.504	2061.948	2095.857	2073.184	2134.136	2123.565
270	2052	2052.955	2064.18	2175.528	1988.078	2105.845	2022.129	2199.371	2017.953
271	2081	2087.326	1919.721	2077.995	1951.182	2138.128	2227.972	2187.318	2018.082

(continued)

Table 17.1 B (continued)

S. No.	Expr RI	Calculated RI							
		Split 1	Split 2	Split 3	Split 4	Split 5	Split 6	Split 7	Split 8
272	2086	2129.915	2196.692	2254.258	2086.695	2190.042	2127.431	2236.952	2103.34

AT is the active training set; PT is a passive training set; CL is the calibration set; VD is the validation set

In this case, we may state that the current model is a mono-parametric correlation of a SMILES-based DCW. The C_0 and C_1 represent regression coefficients [33, 34]. The T and Nepoch denote the Monte Carlo optimization's threshold and the number of epochs, respectively. The T^* and N^* are numeric of T and Nepoch for which the calibration set gives the best statistical result of determination coefficient [35–37].

17.2.3 Optimal Descriptor

In the CORAL programme, the optimal descriptor was derived employing three kinds of descriptors: graph, SMILES and hybrid (graph + SMILES) [34, 38–43]. In this work, the SMILES optimum descriptor of correlation weights (DCW) was used to create QSRR models.

The optimal descriptor DCW based on SMILES was calculated using the following mathematical equation

$$\begin{aligned} \text{SMILES}_{\text{DCW}(T^*N^*)} = & \sum \text{CW}(S_K) + \sum \text{CW}(\text{SS}_K) \\ & + \sum \text{CW}(\text{SSS}_K) + \sum \text{CW}(\text{BOND}) \\ & + \sum \text{CW}(\text{NOSP}) + \sum \text{CW}(\text{HARD}) \\ & + \sum \text{CW}(\text{APP}) + \text{CW}(C_{\text{max}}) + \text{CW}(O) \\ & + \text{CW}(N) + \text{CW}(S) + \text{CW}(=) \end{aligned} \quad (17.2)$$

The description for the depiction used in the above mathematical equation is explained in Table 17.2.

Table 17.2 The detailed description of SMILES attributes

S. No.	SMILES notation	Comments
1	S_k	One SMILES notation or two SMILES notation which cannot be inspected independently
2	SS_k	An amalgamation of two SMILES notations
3	SSS_k	An amalgamation of three SMILES notations
4	BOND	The presence or absence of double ('='), triple ('#') and stereochemical ('@') bonds
5	NOSP	Presence or absence of nitrogen, oxygen, sulphur and phosphorus
6	HARD	Association of BOND, NOSP and HALO
7	APP	Atomic pair proportions of oxygen, nitrogen, sulphur and double bond
8	C_{max}	Contributions of the total number of rings
9	O, N, S and =	Contributions of oxygen, nitrogen, sulphur and double bond

17.2.4 Monte Carlo Optimization

In this study, two target functions, TF_1 ($CII_{\text{weight}} = 0.0$) and TF_2 ($CII_{\text{weight}} = 0.3$), were used to create reliable QSRR models. The statistical outputs of each target function were juxtaposed and analysed.

17.2.4.1 Target Function 1 (TF_1)

The balance of correlation approach was applied to compute the target function 1 (TF_1) and the following mathematical equation is employed to represent it [44, 45].

$$TF_1 = R_{\text{active training}} + R_{\text{passive training}} - |R_{\text{active training}} - R_{\text{passive training}}| \times 0.1 \quad (17.3)$$

Here, R is the correlation coefficient of a specific set.

17.2.4.2 Target Function 2 (TF_2)

The weight of CII was added to TF_1 to compute TF_2 [19, 20, 24], which is expressed by the expression given below.

$$TF_3 = TF_1 + CII_{\text{calibration set}} \times 0.3 \quad (17.4)$$

CII stands for the calibration set's correlation intensity index, and it's determined utilizing the underlying equations.

$$CII = 1 - \sum \Delta R_n^2 > 0 \quad (17.5)$$

$$\Delta R_n^2 = R_n^2 - R^2 \quad (17.6)$$

Here, R^2 stands for determination coefficient of all endpoints and R_n^2 stands for determination coefficient all endpoints excluding n th compound.

17.2.5 Applicability Domain

The application domain (AD), which is the third pillar of the Organization for Economic Cooperation and Development (OECD), is another critical component to include in any built QSAR model [36]. The AD is an imaginary chemical region that includes both model characteristics and expected response. When building a QSAR model, the AD of molecules is used to calculate the degree of uncertainty in the prediction of a particular chemical based on how similar it is to the substances

chosen to make the model. The prediction of a modelled response by QSAR is only valid if the molecule being predicted is inside the AD of the model because it is challenging to predict the entire spectrum of compounds using a single statistical model.

The allotment of SMILES traits in the active training, passive training and calibration sets is utilized to define AD in the QSPR/QSRR models of the CORAL programme [35, 46, 47].

In the present work, VOCs falls in AD if

$$\text{DefectNS} < 2 \times \overline{\text{DefectNS}} \quad (17.7)$$

$$\text{DefectNS} = \sum_{\text{active } A_K} SA_{\text{defect}} \quad (17.8)$$

Here, $\overline{\text{DefectNS}}$ is the average of the statistical defect for the training set.

17.2.6 Validation

The fourth OECD principle highlights the need for information on the efficacy of QSAR models, stating that models should be linked with appropriate goodness-of-fit, robustness (internal performance) and predictability metrics (external performance). Statistical validation methodologies provide a number of “fitness” criteria that QSAR researchers may use to evaluate the performance of various models and avoid models that are either too basic or too sophisticated.

Three methodologies were used in this work to investigate the robustness, reliability and predictive ability of the QSRR models: (i) Internal validation or cross-validation; (ii) External validation; and (iii) Y-scrambling or data randomization. Table 17.3 denotes the equalities for the numerous validation standards (R^2 , CCC, Q^2 , Q^2_{F1} , Q^2_{F2} , Q^2_{F3} , r_m^2 and MAE).

17.2.7 Consensus Modelling

The reliability and effectiveness of the derived QSRR models are assessed by employing certain validation parameters. Researchers desire to improve prediction reliability by minimizing calculated residuals for the developed models. In earlier research, consensus models-which comprised all unique models were found to be more accurately predictive than a specific model. Accordingly, the “*Intelligent Consensus Predictor*” tool developed by Roy et al. was employed to develop consensus models [17, 29, 57]. The following three strategies were typically used to improve the models’ ability to predict outcomes.

Table 17.3 The mathematical equations for the various validation criteria

Validation parameters	References
$R^2 = 1 - \frac{\sum(Y_{\text{obs}} - Y_{\text{prd}})^2}{\sum(Y_{\text{obs}} - \bar{Y})^2}$	[48, 49]
$Q^2 = 1 - \frac{\sum(Y_{\text{prd}} - Y_{\text{obs}})^2}{\sum(Y_{\text{obs}} - \bar{Y}_{\text{train}})^2}$	[49, 50]
$Q_{F1}^2 = 1 - \frac{\sum(Y_{\text{per}(\text{test})} - Y_{\text{obs}(\text{test})})^2}{\sum(Y_{\text{obs}(\text{test})} - \bar{Y}_{\text{train}})^2}$	[51]
$Q_{F2}^2 = 1 - \frac{\sum(Y_{\text{prd}(\text{test})} - Y_{\text{obs}(\text{test})})^2}{\sum(Y_{\text{obs}(\text{test})} - \bar{Y}_{\text{ext}})^2}$	[51, 52]
$Q_{F3}^2 = 1 - \frac{\sum(Y_{\text{prd}(\text{test})} - Y_{\text{obs}(\text{test})})^2/n_{\text{ext}}}{\sum(Y_{\text{obs}(\text{test})} - \bar{Y}_{\text{train}})^2/n_{\text{train}}}$	[51]
$r_m^2 = r^2 \times \left(1 - \sqrt{r^2 - r_0^2}\right)$	[53]
$CCC = \frac{2 \sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2 + \sum(Y - \bar{Y})^2 + n((\bar{X} - \bar{Y})^2)}$	[51, 54]
$MAE = \frac{1}{n} \times \sum Y_{\text{obs}} - Y_{\text{prd}} $	[55, 56]

- (a) **Consensus model 1 (CM1): Average of predictions from all qualified individual models:** It is simply the arithmetic mean of the predicted response scores obtained for a specific sample molecule from all “*N* qualifying individual models (IM)”.
- (b) **Consensus model 2 (CM2): Weighted average predictions (WAPs) from all qualified individual models:** The average for a CM2 is calculated by assigning proportional weightage to authorized models for a given test molecule. Initially, the value of absolute prediction error (AE) for the listed training compounds is assessed for a specific model, henceforth, the result of AE is employed to compute the mean absolute error (MAE_{cv}). The following mathematical relationship is applied to calculate the WAP for a particular query molecule (*k*th compound).

$$\begin{aligned} & \text{WAP}_{\text{Test Object}(k)} \\ &= \frac{[(\text{Pred}_{(k)\text{Model}1} \times w_1) + (\text{Pred}_{(k)\text{Model}2} \times w_2) + \dots + (\text{Pred}_{(k)\text{Model}n} \times w_n)]}{[w_1 + w_2 + \dots + w_n]} \end{aligned} \quad (17.9)$$

Here, weightage (*w*) is computed by the following equation

$$w_1 = \frac{1}{rf_1}, w_2 = \frac{1}{rf_2}, w_3 = \frac{1}{rf_3}, \dots, w_n = \frac{1}{rf_n} \quad (17.10)$$

The following equations are used to compute the relative fraction (rf) for the individual model

$$rf_1 = \frac{MAE_{cv(1)}}{\sum MAE_{CV}}, rf_2 = \frac{MAE_{cv(2)}}{\sum MAE_{CV}}, rf_3 = \frac{MAE_{cv(3)}}{\sum MAE_{CV}}, \dots rf_n = \frac{MAE_{cv(n)}}{\sum MAE_{CV}} \quad (17.11)$$

- (c) Consensus **model 3 (CM3): Best selection of predictions (compound-wise) from individual models:** To predict that specific test set molecule, the best model with the lowest MAE_{cv} is used.

17.3 Results and Discussion

17.3.1 QSRR Modelling and Validation

In the present research, a total of sixteen QSRR models (Eqs. 17.12–17.27) were built using two target functions (TF₁ and TF₂) and the balance of correlation method was applied to obtain unswerving statistical findings. Table 17.2 shows the statistical findings of all QSRR models. To achieve a maximum prediction accuracy of T and Nepoch for all splits, the threshold and Nepoch values were set between 1–10 and 1–25, accordingly. The numeric number for the probe was three while computing the best QSRR model. The numerical value of d_{start} was 0.5 and D_{limit} was 0.1. The weight of CII was 0.0 for TF₁. On the other hand, in the case of TF₂, the weight of CII was used as 0.3. Following the principle that “*QSAR is a random event*,” we made eight random splits to make 16 QSRR models.

The built QSRR models of VOCs based on TF₁ (CII_{weight} = 0.0) for all splits are the following:

$$\begin{aligned} RI &= 379.5804434(\pm 4.9670785) + 87.3071496(\pm 0.4360032) \\ & * DCW(2, 4) \text{ for Split 1} \end{aligned} \quad (17.12)$$

$$\begin{aligned} RI &= 776.1713518(\pm 3.5923744) + 124.9844144(\pm 0.7137551) \\ & * DCW(5, 6) \text{ for Split 2} \end{aligned} \quad (17.13)$$

$$\begin{aligned} RI &= 149.4558572(\pm 13.6894572) + 127.2163725(\pm 1.4957200) \\ & * DCW(7, 7) \text{ for Split 3} \end{aligned} \quad (17.14)$$

$$\begin{aligned} RI &= 128.62300399(\pm 5.9858047) + 103.1167439(\pm 0.5663511) \\ & * DCW(7, 13) \text{ for Split 4} \end{aligned} \quad (17.15)$$

$$\begin{aligned} RI &= -126.5460843(\pm 8.2693132) + 43.5125747(\pm 0.2411775) \\ & * DCW(1, 6) \text{ for Split 5} \end{aligned} \quad (17.16)$$

$$\begin{aligned} \text{RI} &= 579.4274718(\pm 3.9208238) + 82.0507869(\pm 0.5053360) \\ & * \text{DCW}(8, 16) \text{for Split 6} \end{aligned} \quad (17.17)$$

$$\begin{aligned} \text{RI} &= 258.1293444(\pm 5.1553852) + 136.5706989(\pm 0.6949786) \\ & * \text{DCW}(3, 8) \text{for Split 7} \end{aligned} \quad (17.18)$$

$$\begin{aligned} \text{RI} &= 218.2068362(\pm 4.9668148) + 84.6015056(\pm 0.3632482) \\ & * \text{DCW}(6, 4) \text{for Split 8} \end{aligned} \quad (17.19)$$

The built QSRR models of VOCs based on TF_2 ($\text{CII}_{\text{weight}} = 0.3$) for all splits are the following:

$$\begin{aligned} \text{RI} &= 398.8188624(\pm 5.2141795) + 65.0057648(\pm 0.3700230) \\ & * \text{DCW}(2, 4) \text{for Split 1} \end{aligned} \quad (17.20)$$

$$\begin{aligned} \text{RI} &= 234.2735057(\pm 5.6567674) + 69.6995549(\pm 0.3358152) \\ & * \text{DCW}(1, 14) \text{for Split 2} \end{aligned} \quad (17.21)$$

$$\begin{aligned} \text{RI} &= -49.0157479(\pm 3.0566361) + 81.2090986(\pm 0.1979594) \\ & * \text{DCW}(1, 15) \text{for Split 3} \end{aligned} \quad (17.22)$$

$$\begin{aligned} \text{RI} &= -516.7046231(\pm 6.4384415) + 98.7363085(\pm 0.4127289) \\ & * \text{DCW}(3, 19) \text{for Split 4} \end{aligned} \quad (17.23)$$

$$\begin{aligned} \text{RI} &= -119.8129213(\pm 8.1464342) + 121.2288111(\pm 0.6665878) \\ & * \text{DCW}(1, 15) \text{for Split 5} \end{aligned} \quad (17.24)$$

$$\begin{aligned} \text{RI} &= 586.7821685(\pm 3.5348059) + 104.8560860(\pm 0.5358107) \\ & * \text{DCW}(2, 13) \text{for Split 6} \end{aligned} \quad (17.25)$$

$$\begin{aligned} \text{RI} &= 135.1254654(\pm 5.1240395) + 101.4684828(\pm 0.4575822) \\ & * \text{DCW}(1, 16) \text{for Split 7} \end{aligned} \quad (17.26)$$

$$\begin{aligned} \text{RI} &= 362.8926110(\pm 3.7981415) + 93.2377006(\pm 0.4038385) \\ & * \text{DCW}(3, 10) \text{for Split 8} \end{aligned} \quad (17.27)$$

Various statistical criteria (R^2 , CCC, CII, Q^2 , Q^2F_1 , Q^2F_2 , Q^2F_3 , s , MAE, F, RMSE, R_m^2 , ΔR_m^2 , $C_{R_p^2}$ and Y-test) were examined in order to assess the models' robustness and predictability. Roy et al. gave statistical parameters (R_m^2 and ΔR_m^2) to in-depth examine the external predictability of a QSAR/QSPR model and defined that R_m^2 is the most strict parameter of external validation [58]. Golbraikh et al. defined the acceptable range of important statistical metrics ($R^2 > 0.6$ and $Q^2 > 0.5$) to validate the robustness of QSAR/QSPR models [49]. Chirico et al. discussed the significance of the Concordance Correlation Coefficient (CCC) as a complementary, or alternative, more sensible measure of a QSAR model [51]. All statistical attributes of proposed models developed by both target functions (TF₁ and TF₂) are listed in Table 17.4A, B and all of the models meet the required criteria of each parameter. The coefficient of determination derived by TF₂ for the validation set of split 4 has the maximum statistical result ($R_{\text{validation}}^2 = 0.9532$), hence it is accepted as the winner model. The developed QSRR models with the second target function (TF₂-optimization) were more robust with better predictive ability. For all splits, it was observed that the numerical value of R^2 calculated by TF₂ for calibration and validation was greater than the R^2 calculated by TF₁.

The applicability domain (AD) is the third OECD guideline, and it describes how a specific compound is relevant to the database employed to build a QSPR model. AD is deployed to identify the outlier in the SMILES-based QSAR model. The number of outliers present in the validation set of QSRR models developed by TF₁ was 4, 3, 3, 1, 3, 2, 3 and 3 for the splits 1, 2, 3, 4, 5, 6, 7 and 8, respectively (see supporting information TF₁). But, in the case of the models built by TF₂, the number of outliers for the validation set were 6, 7, 4, 3, 3, 9, 5 and 3 for the splits 1, 2, 3, 4, 5, 6, 7 and 8, respectively (see supporting information TF₂).

To classify statistical results as "improved" or "unimproved," the following conditions are used (Table 17.4A, B) [17, 57, 59–61].

$$\text{if } X_{\text{Cib}}[\text{TF}_2] > X_{\text{Cib}}[\text{TF}_1] \text{ and } R_{\text{valid}}^2[\text{TF}_2] > R_{\text{valid}}^2[\text{TF}_1] \quad (17.28)$$

(Then, the validation parameters were labelled as "improved")

$$\text{if } X_{\text{Cib}}[\text{TF}_2] < X_{\text{Cib}}[\text{TF}_1] \text{ and } R_{\text{valid}}^2[\text{TF}_2] < R_{\text{valid}}^2[\text{TF}_1] \quad (17.29)$$

and

$$\text{if } X_{\text{Cib}}[\text{TF}_2] \{ X_{\text{Cib}}[\text{TF}_1] \text{ and } R_{\text{valid}}^2[\text{TF}_2] \} R_{\text{valid}}^2[\text{TF}_1] \quad (17.30)$$

(Then, the validation parameters were labelled as "unimproved").

where $X_{\text{Cib}}[\text{TF}_2]$ and $X_{\text{Cib}}[\text{TF}_1]$ are the statistical results of different statistical standards (R^2 , CCC, IIC, Q^2 , Q^2F_1 , Q^2F_2 , Q^2F_3 , s , MAE, F, RMSE, Avg R_m^2 , ΔR_m^2 , $C_{R_p^2}$ and Y-test) of the calibration set. All statistical standards of QSRR models established for splits 3, 6 and 7 were improved by CII. The index of ideality correlation (IIC) was not enhanced in splits 2 and 1. The numerical value of R_r^2 was not improved

Table 17.4 A The statistical parameters of constructed QSAR models for full data set encompassing active training set (AT), passive training set (PT), calibration set (CL) and validation set (VD) of VOCs using TF₁ and TF₂

Split	Target function	Set	<i>n</i>	<i>R</i> ²	CCC	IIC	CII	Q ²	Q ² F ₁	Q ² F ₂	Q ² F ₃		
1	TF ₁	AT	34	0.9854	0.9926	0.9927	0.9895	0.9829					
		PT	34	0.9622	0.9679	0.2909	0.9723	0.9554					
		CL	34	0.9132	0.9436	0.4535	0.9493	0.9005	0.8947	0.8945	0.8925		
		VD	34	0.8857	0.9322	0.7961	0.9300	0.8612					
	TF ₂	AT	34	0.9870	0.9935	0.8831	0.9902	0.9842					
		PT	34	0.9560	0.9730	0.3699	0.9655	0.9473					
		CL	34	0.9605	0.9739	0.3215	0.9782	0.9523	0.9510	0.9509	0.9500		
		VD	34	0.9401	0.9679	0.7096	0.9570	0.9328					
	Comments				Improved	Improved	Unimproved	Improved	Improved	Improved	Improved	Improved	
	2	TF ₁	AT	34	0.9703	0.9849	0.6895	0.9807	0.9665				
			PT	34	0.9813	0.9885	0.4785	0.9860	0.9768				
			CL	34	0.9449	0.9697	0.6427	0.9609	0.9379	0.9378	0.9373	0.9395	
VD			34	0.8191	0.8951	0.6145	0.8637	0.8013					
TF ₂		AT	34	0.9888	0.9944	0.8839	0.9914	0.9866					
		PT	34	0.9886	0.9920	0.7334	0.9911	0.9865					
		CL	34	0.9810	0.9883	0.5466	0.9880	0.9781	0.9767	0.9765	0.9773		
		VD	34	0.9490	0.9664	0.9341	0.9662	0.9422					
Comments				Improved	Improved	Unimproved	Improved	Improved	Improved	Improved	Improved		
3		TF ₁	AT	34	0.9226	0.9597	0.9605	0.9544	0.9095				
			PT	34	0.9620	0.9744	0.5432	0.9720	0.9574				
			CL	34	0.9366	0.9657	0.5624	0.9600	0.9243	0.9323	0.9318	0.9321	

(continued)

Table 17.4 A (continued)

Split	Target function	Set	<i>n</i>	<i>R</i> ²	CCC	IIC	CII	<i>Q</i> ²	<i>Q</i> ² <i>F</i> ₁	<i>Q</i> ² <i>F</i> ₂	<i>Q</i> ² <i>F</i> ₃		
4	TF ₂	VD	34	0.8700	0.9326	0.7476	0.9102	0.8496					
		AT	34	0.9926	0.9963	0.8856	0.9946	0.9918					
		PT	34	0.9931	0.9896	0.8133	0.9953	0.9922					
		CL	34	0.9740	0.9868	0.7269	0.9891	0.9686	0.9734	0.9732	0.9733		
		VD	34	0.9305	0.9629	0.4594	0.9492	0.9220					
	Comments				Improved	Improved	Improved	Improved	Improved	Improved	Improved	Improved	
		TF ₁	AT	34	0.9648	0.9821	0.6081	0.9718	0.9607				
			PT	34	0.9612	0.9709	0.5698	0.9767	0.9539				
			CL	34	0.8546	0.9171	0.5510	0.8923	0.8198	0.8390	0.8385	0.8353	
		VD	34	0.9376	0.9666	0.7875	0.9559	0.9267					
TF ₂	AT	34	0.9781	0.9889	0.8791	0.9817	0.9756						
	PT	34	0.9730	0.9700	0.8524	0.9772	0.9679						
	CL	34	0.9574	0.9732	0.4998	0.9726	0.9497	0.9501	0.9499	0.9489			
	VD	34	0.9532	0.9732	0.5302	0.9660	0.9462						
5	Comments			Improved	Improved	Improved	Improved	Improved	Improved	Improved	Improved		
		TF ₁	AT	34	0.9857	0.9928	0.6146	0.9866	0.9829				
			PT	34	0.9861	0.9880	0.7013	0.9900	0.9846				
			CL	34	0.9155	0.9540	0.6036	0.9389	0.9020	0.9113	0.9106	0.9060	
		VD	34	0.9150	0.9349	0.4293	0.9364	0.9057					
	TF ₂	AT	34	0.9864	0.9931	0.6952	0.9877	0.9836					
		PT	34	0.9861	0.9894	0.9303	0.9904	0.9844					

(continued)

Table 17.4 A (continued)

Split	Target function	Set	<i>n</i>	<i>R</i> ²	CCC	IIC	CII	<i>Q</i> ²	<i>Q</i> ² <i>F</i> ₁	<i>Q</i> ² <i>F</i> ₂	<i>Q</i> ² <i>F</i> ₃	
6		CL	34	0.9782	0.9872	0.8200	0.9908	0.9739	0.9760	0.9758	0.9745	
		VD	34	0.9496	0.9673	0.2686	0.9603	0.9451				
	Comments			Improved	Improved	Improved	Improved	Improved	Improved	Improved	Improved	
	TF ₁	AT	34	0.9630	0.9812	0.7747	0.9750	0.9584				
		Pass	34	0.9717	0.9825	0.7034	0.9793	0.9666				
	TF ₂	CL	34	0.9082	0.9462	0.4468	0.9256	0.8969	0.8927	0.8924	0.8899	
		VD	34	0.8175	0.8929	0.7413	0.8616	0.7696				
		AT	34	0.9781	0.9889	0.8791	0.9853	0.9750				
		PT	34	0.9757	0.9851	0.8787	0.9809	0.9716				
		CL	34	0.9798	0.9865	0.5347	0.9885	0.9760	0.9716	0.9715	0.9708	
		VD	34	0.8928	0.9303	0.8512	0.9160	0.8812				
	Comments			Improved	Improved	Improved	Improved	Improved	Improved	Improved	Improved	
7	TF ₁	AT	34	0.9882	0.9941	0.5422	0.9919	0.9858				
		PT	34	0.9879	0.9898	0.9751	0.9911	0.9864				
	TF ₂	CL	34	0.8622	0.9147	0.5301	0.9225	0.8306	0.7953	0.7947	0.8043	
		VD	34	0.9114	0.9387	0.7223	0.9448	0.8969				
	TF ₂	AT	34	0.9934	0.9967	0.7869	0.9947	0.9917				
		PT	34	0.9948	0.9919	0.9311	0.9957	0.9941				
	Comments	CL	34	0.9714	0.9695	0.8370	0.9901	0.9648	0.9260	0.9258	0.9293	
		VD	34	0.9252	0.9484	0.8248	0.9441	0.9153				
				Improved	Improved	Improved	Improved	Improved	Improved	Improved	Improved	Improved

(continued)

Table 17.4 B The statistical parameters of constructed QSAR models for full data set encompassing active training set (AT), passive training set (PT), calibration set (CL) and validation set (VD) of VOCs using TF1 and TF2

Split	Target function	Set	n	RMSE	MAE	F	R_t^2	$C_{R_p}^2$	Avg. R_m^2	Delta R_m^2	
1	TF ₁	AT	34	48.2	34.0	2155	0.0298	0.9704			
		PT	34	107.1	68.8	814	0.0308	0.9467			
		CL	34	131.4	90.2	337	0.0519	0.8869	0.8744	0.0557	
		VD	34	155.9	98.2	248	0.0379	0.8665	0.7724	0.1121	
		TF ₂	AT	34	45.4	31.6	2434	0.0543	0.9595		
			PT	34	96.2	54.3	695	0.0157	0.9481		
	CL		34	89.6	62.5	778	0.0554	0.9324	0.9248	0.0376	
	VD		34	102.8	72.3	502	0.0335	0.9232	0.8864	0.0571	
	Comments				Improved	Improved	Unimproved	Improved	Improved	Improved	
						Improved	Improved	Improved	Improved	Improved	Improved
	2	TF ₁	AT	34	69.4	47.8	1047	0.0369	0.9517		
			PT	34	59.9	41.9	1678	0.0186	0.9719		
CL			34	98.6	67.7	549	0.0537	0.9177	0.9198	0.0184	
VD			34	172.9	111.5	145	0.0426	0.7975	0.7044	0.1636	
TF ₂			AT	34	42.7	25.2	2821	0.0609	0.9579		
			PT	34	48.8	33.8	2776	0.0273	0.9749		
		CL	34	60.4	44.2	1649	0.0072	0.9773	0.9606	0.0183	
		VD	34	98.9	71.7	596	0.0407	0.9284	0.8330	0.0602	
		Comments				Improved	Improved	Improved	Improved	Improved	Improved
						Improved	Improved	Improved	Improved	Improved	Improved
3		TF ₁	AT	34	112.6	81.5	381	0.0372	0.9037		
			PT	34	90.0	66.5	810	0.0324	0.9457		
	CL		34	104.0	69.0	473	0.0306	0.9212	0.9077	0.0544	

(continued)

Table 17.4 B (continued)

Split	Target function	Set	<i>n</i>	RMSE	MAE	F	R_r^2	C_{Rp}^2	Avg. R_m^2	Delta R_m^2	
4	TF ₂	VD	34	147.5	100.5	214	0.0379	0.8508	0.8147	0.0026	
		AT	34	34.7	25.0	4316	0.0184	0.9834			
		PT	34	60.0	44.2	4603	0.0111	0.9875			
		CL	34	65.2	51.6	1198	0.0206	0.9636	0.9616	0.0242	
		VD	34	113.4	87.1	429	0.0388	0.9109	0.8772	0.0649	
	Comments				Improved	Improved	Improved	Improved	Improved	Improved	Improved
		TF ₁	AT	34	73.8	52.2	878	0.0362	0.9466		
			PT	34	92.0	62.4	793	0.0311	0.9455		
			CL	34	162.0	93.4	188	0.0214	0.8438	0.7935	0.0118
		VD	34	100.5	67.3	481	0.0364	0.9192	0.9055	0.0573	
TF ₂	AT	34	58.2	37.1	1431	0.0289	0.9636				
5	Comments			Improved	Improved	Improved	Improved	Improved	Improved	Unimproved	
		TF ₁	AT	34	47.3	20.7	2212	0.0364	0.9674		
			PT	34	64.6	46.3	2276	0.0615	0.9549		
			CL	34	121.7	82.3	347	0.0238	0.9035	0.8782	0.0130
		VD	34	157.3	97.4	345	0.0458	0.8918	0.7299	0.1096	
	TF ₂	AT	34	46.3	26.2	2315	0.0478	0.9622			
	PT	34	60.3	40.9	2269	0.0409	0.9654				

(continued)

Table 17.4 B (continued)

Split	Target function	Set	n	RMSE	MAE	F	R_p^2	$C_{R_p}^2$	Avg. R_m^2	Delta R_m^2	
6		CL	34	63.3	43.5	1435	0.0200	0.9681	0.9255	0.0230	
		VD	34	105.7	65.5	604	0.0449	0.9269	0.8542	0.0548	
	Comments			Improved	Improved	Improved	Improved	Improved	Improved	Unimproved	
	TF ₁	AT	34	76.0	60.4	833	0.0352	0.9453			
		Pass	34	71.2	48.3	1097	0.0290	0.9571			
	TF ₂	CL	34	131.4	74.2	317	0.0420	0.8870	0.8674	0.0587	
		VD	34	175.3	107.4	143	0.0293	0.8027	0.6881	0.1685	
		AT	34	58.5	43.6	1426	0.0158	0.9701			
		PT	34	66.5	48.7	1287	0.0702	0.9400			
	Comments	CL	34	67.6	46.5	1552	0.0092	0.9752	0.9141	0.0230	
VD		34	140.6	90.7	267	0.0322	0.8766	0.7689	0.1098		
			Improved	Improved	Improved	Improved	Improved	Improved	Improved	Improved	
			Improved	Improved	Improved	Improved	Improved	Improved	Improved	Improved	
7	TF ₁	AT	34	44.1	29.7	2677	0.0174	0.9795			
		PT	34	59.8	48.3	2618	0.0250	0.9753			
	TF ₂	CL	34	178.9	105.0	200	0.0441	0.8398	0.7135	0.1394	
		VD	34	152.9	106.3	329	0.0400	0.8912	0.7622	0.1008	
		AT	34	32.9	18.8	4848	0.0172	0.9848			
		PT	34	53.9	43.2	6074	0.0195	0.9849			
	Comments	CL	34	107.5	73.9	1086	0.0211	0.9608	0.8153	0.0496	
		VD	34	139.1	100.5	396	0.0390	0.9055	0.7769	0.0888	
				Improved	Improved	Improved	Improved	Improved	Improved	Improved	Improved
				Improved	Improved	Improved	Improved	Improved	Improved	Improved	Improved

(continued)

Table 17.4 B (continued)

Split	Target function	Set	n	RMSE	MAE	F	R_r^2	$C_{R_p}^2$	Avg. R_m^2	Delta R_m^2
8	TF ₁	AT	34	67.1	47.5	1107	0.0103	0.9667		
		PT	34	68.3	50.7	1344	0.0368	0.9582		
		CL	34	108.4	84.2	468	0.0388	0.9164	0.9067	0.0523
		VD	34	186.6	119.7	185	0.0355	0.8344	0.7202	0.1419
	TF ₂	AT	34	45.4	36.5	2455	0.0210	0.9766		
		PT	34	59.0	39.9	2611	0.0395	0.9679		
		CL	34	79.9	57.5	825	0.0460	0.9394	0.9454	0.0126
		VD	34	141.4	75.3	271	0.0374	0.875400291	0.8186	0.0975
Comments										
				Improved	Improved	Improved	Unimproved	Improved	Improved	Improved

n = Number of datum in set; R^2 = Correlation coefficient; CCC= Concordance correlation coefficient; IIC = Index of ideality of correlation; CII = Correlation intensity index; Q^2 = leave-one-out cross validated correlation coefficient, Q^2F_1 , Q^2F_2 , Q^2F_3 are validation matrices; RMSE = Root mean square error; MAE = Mean absolute error; F = Fischer's ratio

in the case of splits 1 and 8. The numerical results of ΔR_m^2 was not improved by the CII for splits 4 and 5. As a result, it is logical to conclude that the correlation intensity index (TF_2) improves the predictive potential of developed QSRR models for the retention indices of 136 volatile organic compounds (VOC) detected in the headspace of rice.

17.3.2 Mechanistic Interpretation

The structural attributes acquired from SMILES were categorized into two primary classes based on the numerical value of correlation weights (CW): reliable promoters (promoters of increase and decrease) and unreliable promoters. The reliable promoters were classified as the promoter of increase if these had consistent positive CW numerical values in all runs/in three or more splits of Monte Carlo optimization, otherwise, these were classified as the promoter of decrease (negative CW numerical values in all runs). The structural attributes were known as unreliable promoters if both negative and positive numerical values for all runs of Monte Carlo optimization were obtained. These promoters were used to provide useful insight into the mechanistic analysis of the QSRR models generated by CORAL software's inbuilt Monte Carlo algorithm. Table 17.5 provides a list of reliable promoters derived from four different splits (splits 1, 2, 3 and 4). The structural attributes (SA_k) such as the presence of sulphur, aliphatic carbon, aromatic carbon aliphatic, cyclic ring, branching and alkene were identified as reliable promoters of increase. The structural attributes such as the absence of sulphur and oxygen were identified as reliable negative promoters.

17.3.3 Consensus Modelling

The scientific study claims that the adoption of an “*intelligent consensus predictor tool*” and consensus modelling increased the predictability of QSAR/QSPR models [29, 57, 62]. To get new training and test sets for consensus modelling, the allocation method of split 4 was employed. The specific information on the results of consensus modelling and individual models is described in Table 17.6. The “Consensus Model 2: Weighted average predictions from “*qualified*” *Individual models*” was the winner model based on MAE (95%; test) [63]. The prediction ability of each model was rated “*good*” using the MAE-based standard. The Dixon-Q test and the applicability domain were employed. The value of threshold (k) and Euclidean Distance cut-off were taken as 3.0 and 0.3, respectively.

The numerical value of R^2 determined from “Consensus Model 2: Weighted average predictions from “*qualified*” *Individual models*” was 0.9864 (Fig. 17.1). The result of consensus modelling was better than the individual modelling. As a result, it can be stated that consensus modelling may be used to predict the retention

Table 17.5 List of reliable promoters derived from four different splits

S. No.	SA _k	CW(SA _k)				Description
		Split 1	Split 2	Split 3	Split 4	
<i>A</i>						
<i>Promoters of endpoint Increase</i>						
1.	\$00,001,000,000	0.0596	2.25613	0.89651	0.44436	Presence of oxygen
2.	< = > .0000...	0.02998	0.1057	0.59068	0.6655	Absence of double bond
3.	1...c...(...	0.39136	0.59142	0.43469	0.59046	Branching on the first aromatic ring
4.	2...	0.41471	0.96504	0.75866	1.22903	Presence of two ring
5.	c...(...	0.17889	0.35486	0.51189	0.00123	Aromatic carbon with branching
6.	C...(...C...	0.39661	0.15471	0.56535	0.28287	Presence of two aliphatic carbon with branching
7.	C.../...	0.12037	0.52454	0.16729	0.3574	Aliphatic carbon with cis/trans bond
8.	C... = ...C...	1.03587	0.32746	0.0524	0.13634	Two aliphatic carbon joined by a double bond
9.	c...1...	0.90253	0.35094	0.16078	0.55609	Presence of one aromatic carbon
10.	c...c...(...	0.14837	0.92669	0.54476	0.30209	Sequential combination of two aromatic carbon followed by branching
11.	C...C...	0.27372	1.18971	0.01506	0.47043	Combination of two aliphatic carbon
12.	c...c...	0.24999	0.39362	0.44382	0.10986	Combination of two aromatic carbon
13.	C...c...1...	1.09246	0.90312	0.09741	0.04667	Presence of aliphatic carbon with aromatic carbon on the first ring
14.	C...C...C...	0.78199	0.08293	0.9807	0.45479	Aliphatic chain of three carbon atom

(continued)

Table 17.5 (continued)

S. No.	SA _k	CW(SA _k)				Description
		Split 1	Split 2	Split 3	Split 4	
15.	Cmax 0001...	0.87504	1.15809	0.24905	1.48948	Presence of maximum one ring
16.	O...(C...	0.2674	0.19941	0.33266	0.53041	Sequential combination of aliphatic carbon, branching and aliphatic carbon
17.	O...	0.05484	0.55447	0.35917	0.33709	Presence of oxygen atom
18.	O... = ...C...	1.61775	1.12829	0.05799	0.24476	Presence of aliphatic oxygen, double bond and carbon
<i>B</i>	<i>Promoters of endpoint Decrease</i>					
1.	\$00,000,000,000	- 1.93441	- 1.21227	- 0.8947	- 0.9151	Absence of HARD: Super-attribute of SMILES
2.	< = > .00001...	- 1.15262	- 0.2895	- 0.10261	- 0.35307	Presence of one double bond
3.	< S > .0.0000...	- 1.67041	- 0.0004	- 0.45945	- 2.56956	Absence of sulphur
4.	C...(C...	- 0.69041	- 0.2988	- 0.40102	- 0.56006	Presence of aliphatic carbon with two branching
5.	Cmax 0000...	- 0.23134	- 1.10586	- 0.72821	- 0.58524	Absence of ring
6.	NOSP00000000	- 0.85768	- 0.38733	- 0.01996	- 1.35213	Absence of nitrogen, oxygen, sulphur and phosphorous

index (RI) of 136 volatile organic compounds (VOCs) detected in the headspace of rice utilizing a Divinylbenzene-Carboxen-Polydimethylsiloxane system (DVB-CAR-PDMS) fibre in the solid-phase micro extraction-gas chromatography-mass spectrometry (SPME-GC-MS) analysis.

Table 17.6 Results of consensus modelling

Type of model	IM1	IM2	IM3	IM4	IM5	IM6	IM7	IM8	CM0	CM1	CM2	CM3
Number of compounds	34	34	34	34	34	34	34	34	34	34	34	34
R ²	0.9870	0.9886	0.9740	0.9532	0.9770	0.9736	0.9745	0.9751	0.9855	0.9855	0.9864	0.9871
Q ² F ₁ (100%)	0.9816	0.9863	0.9722	0.9512	0.9768	0.9726	0.9743	0.9739	0.9851	0.9851	0.9860	0.9870
Q ² F ₂ (100%)	0.9815	0.9862	0.9721	0.9510	0.9767	0.9725	0.9742	0.9738	0.9851	0.9851	0.9859	0.9869
Q ² F ₃ (100%)	0.9817	0.9864	0.9724	0.9515	0.9770	0.9728	0.9745	0.9741	0.9852	0.9852	0.9861	0.9871
CCC(100%)	0.9903	0.9928	0.9852	0.9759	0.9880	0.9863	0.9869	0.9867	0.9923	0.9923	0.9927	0.9933
Avg R _m ² (100%)	0.9177	0.9631	0.9253	0.9298	0.9500	0.9562	0.9533	0.9456	0.9584	0.9584	0.9596	0.9660
Delta R _m ² (100%)	0.0163	0.0110	0.0257	0.0417	0.0214	0.0239	0.0234	0.0232	0.0138	0.0138	0.0130	0.0121
MAE(100%)	37.0134	34.2770	45.8479	60.4356	39.1129	47.3286	43.2513	39.4674	31.4957	31.4957	30.7461	31.2072
MAE(95%)	30.0795	28.5185	37.3947	50.3881	29.7930	38.5682	33.9966	28.1912	24.4652	24.4652	23.7780	25.5798
PRESS(100%)	96,713.5	72,274.4	146,144.3	256,654.9	121,923.5	144,153.1	135,294.5	137,162.6	78,144.1	78,144.1	73,868.5	68,537.8
PRESS(95%)	40,223.9	38,127.6	80,460.6	158,269.6	49,111.6	72,351.2	58,468.3	40,205.0	29,699.5	29,699.5	28,051.3	36,213.5
SDEP(100%)	53.3340	46.1055	65.5619	86.8831	59.8831	65.1137	63.0813	63.5153	47.9412	47.9412	46.6112	44.8979
SDEP(95%)	35.4541	34.5179	50.1437	70.3273	39.1757	47.5497	42.7450	35.4458	30.4649	30.4649	29.6075	33.6403

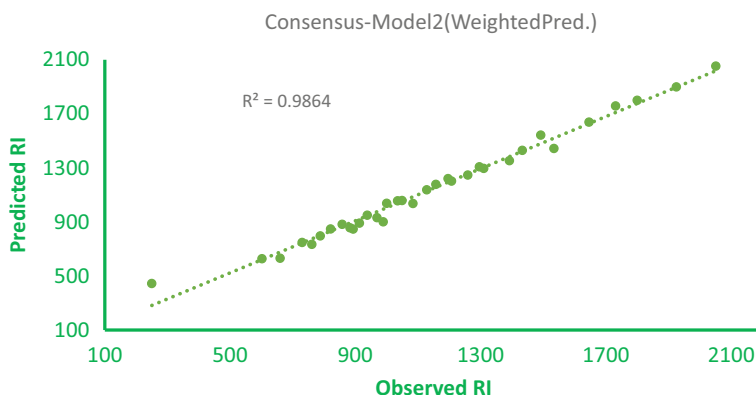


Fig. 17.1 The correlation between the observed RI and the predicted RI computed by CM2

17.4 Conclusions

The retention indices of 136 volatile organic compounds identified by the DVB-CAR-PDMS fibre in the headspace of rice are explained and predicted by a SMILES-based QSRR modelling using the CORAL software. The newly introduced statistical parameter “*Correlation Intensity Index (CII)*” is also applied. The results of validation parameters are improved by using the CII. Two target functions, TF_1 ($W_{CII} = 0$) and TF_2 ($W_{CII} = 0.3$) are implemented to make 16 QSRR models from eight random splits employing the balance of correlation scheme. The result of $R^2_{\text{validation}} = 0.9532$ of TF_2 (Split 4) is obtained better than that of the $R^2_{\text{validation}}$ of the other splits, for this reason, it is accredited as a prominent model. The mechanistic interpretation is also done by computing the reliable structural attributes. The structural attributes (SA_K) such as the presence of sulphur, aliphatic carbon, aromatic carbon aliphatic, cyclic ring, branching and alkene were identified as reliable promoters of increase. The structural attributes such as the absence of sulphur and oxygen were identified as reliable negative promoters. Finally, a consensus model is built using the allocation method of split 4 and the “Consensus Model 2: Weighted average predictions from “*qualified*” *Individual models*” is found best model based on MAE (95%; test). The numerical result of R^2 of the test set for the CM2 model is found 0.9864. Hence, it can be concluded that the present QSRR methodology can be applied to predict the retention indices of 136 volatile organic compounds.

Acknowledgements The authors are also thankful to their respective universities for providing the infrastructure.

Disclosure Statement No potential conflict of interest was reported by the authors.

Data Availability The processed data needed to reproduce these findings are provided in the manuscript.

References

1. Ramtekey V, Cherukuri S, Modha KG, Kumar A, Kethineni UB, Pal G, Singh AN, Kumar S (2021) *Rev Anal Chem* 40(1):272–292. <https://doi.org/10.1515/revac-2021-0137>
2. Ma Z-H, Wang Y-B, Cheng H-T, Zhang G-C, Lyu W-Y (2020) *Food Chem* 311:125896. <https://doi.org/10.1016/j.foodchem.2019.125896>
3. Liu K-l, Zheng J-b, Chen F-s (2017) *LWT - Food Sci Tech* 82:429–436. <https://doi.org/10.1016/j.lwt.2017.04.067>
4. Jia M, Wang X, Liu J, Wang R, Wang A, Strappe P, Shang W, Zhou Z (2022) *Food Chem* 371:131119. <https://doi.org/10.1016/j.foodchem.2021.131119>
5. Kasote D, Singh VK, Bollinedi H, Singh AK, Sreenivasulu N, Regina A (2021) *Foods* 10(8):1917. <https://doi.org/10.3390/foods10081917>
6. Jinakot I, Jirapakkul W (2019) *J Nut Sci Vitaminol* 65(Supplement):S231–S234. <https://doi.org/10.3177/jnsv.65.S231>
7. Mahattanatawee K, Rouseff RL (2014) *Food Chem* 154:1–6. <https://doi.org/10.1016/j.foodchem.2013.12.105>
8. Kalisznan R (2007) *Chem Rev* 107(7):3212–3246. <https://doi.org/10.1021/cr068412z>
9. Rojas C, Tripaldi P, Pérez-González A, Duchowicz PR, Pis Diez R (2018) *J Cereal Sci* 79:303–310. <https://doi.org/10.1016/j.jcs.2017.11.004>
10. Kumar A, Kumar P (2020) *J Mol Liq* 318:114055. <https://doi.org/10.1016/j.molliq.2020.114055>
11. Lotfi S, Ahmadi S, Kumar P (2021) *J Mol Liq* 338:116465. <https://doi.org/10.1016/j.molliq.2021.116465>
12. Lotfi S, Ahmadi S, Kumar P (2021) *RSC Adv* 11(54):33849–33857. <https://doi.org/10.1039/d1ra06861j>
13. Kumar A, Bagri K, Nimbhal M, Kumar P (2021) *J Biomol Struct Dyn* 39(18):7181–7193. <https://doi.org/10.1080/07391102.2020.1806111>
14. Kumar A, Sindhu J, Kumar P (2021) *J Biomol Struct Dyn* 39(14):5014–5025. <https://doi.org/10.1080/07391102.2020.1784286>
15. Ahmadi S, Lotfi S, Afshari S, Kumar P, Ghasemi E (2021) *SAR QSAR Environ Res* 32(12):1013–1031. <https://doi.org/10.1080/1062936X.2021.2003429>
16. Kumar P, Kumar A, Singh D (2022) *Environ Toxicol Pharmacol* 93:103893. <https://doi.org/10.1016/j.etap.2022.103893>
17. Kumar P, Kumar A, Lal S, Singh D, Lotfi S, Ahmadi S (2022) *J Mol Struct* 1265:133437. <https://doi.org/10.1016/j.molstruc.2022.133437>
18. Jafari K, Fatemi MH, Toropova AP, Toropov AA (2022) *Chemometr Intell Lab Syst* 222:104500. <https://doi.org/10.1016/j.chemolab.2022.104500>
19. Kumar P, Kumar A (2021) *J Mol Struct* 1246:131205. <https://doi.org/10.1016/j.molstruc.2021.131205>
20. Toropova AP, Toropov AA (2020) *Fuller Nanotub Carbon Nanostruct* 28(11):900–906. <https://doi.org/10.1080/1536383X.2020.1779705>
21. Toropov AA, Toropova AP (2020) *Sci Total Environ* 737:139720. <https://doi.org/10.1016/j.scitotenv.2020.139720>
22. Toropov AA, Sizochenko N, Toropova AP, Leszczynska D, Leszczynski J (2020) *J Mol Liq* 317:113929. <https://doi.org/10.1016/j.molliq.2020.113929>
23. Jafari K, Fatemi MH, Toropova AP, Toropov AA (2020) *Chem Phy Lett* 754:137614. <https://doi.org/10.1016/j.cplett.2020.137614>
24. Ahmadi S, Toropova AP, Toropov AA (2020) *Nanotoxicol* 14(8):1118–1126. <https://doi.org/10.1080/17435390.2020.1808252>
25. Papa E, van der Wal L, Arnot JA, Gramatica P (2014) *Sci Total Environ* 470–471:1040–1046. <https://doi.org/10.1016/j.scitotenv.2013.10.068>
26. Papa E, Gramatica P (2010) *Green Chem* 12(5):836–843. <https://doi.org/10.1039/B923843C>
27. Zhu H, Tropsha A, Fourches D, Varnek A, Papa E, Gramatica P, Öberg T, Dao P, Cherkasov A, Tetko IV (2008) *J Chem Inf Model* 48(4):766–784. <https://doi.org/10.1021/ci700443v>

28. Kumar P, Kumar A (2021) *Nanotoxicol* 15(9):1199–1214. <https://doi.org/10.1080/17435390.2021.2008039>
29. Roy K, Ambure P, Kar S, Ojha PK (2018) *J Chemom* 32(4):e2992. <https://doi.org/10.1002/cem.2992>
30. Marvin-Sketch-v.14.11.17.0 (2014). ChemAxon, XchemAxon KFT. Budapest, Hungary
31. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) *J Cheminform* 3(1):33. <https://doi.org/10.1186/1758-2946-3-33>
32. Kumar P, Kumar A (2020) *Chemometr Intel Lab Syst* 200:103982. <https://doi.org/10.1016/j.chemolab.2020.103982>
33. Duhan M, Sindhu J, Kumar P, Devi M, Singh R, Kumar R, Lal S, Kumar A, Kumar S, Hussain K (2022) *J Biomol Struct Dyn* 40(11):4933–4953. <https://doi.org/10.1080/07391102.2020.1863861>
34. Ahmadi S, Lotfi S, Kumar P (2020) *SAR QSAR Environ Res* 31(12):935–950. <https://doi.org/10.1080/1062936X.2020.1842495>
35. Duhan M, Singh R, Devi M, Sindhu J, Bhatia R, Kumar A, Kumar P (2021) *J Biomol Struct Dyn* 39(1):91–107. <https://doi.org/10.1080/07391102.2019.1704885>
36. Kumar P, Kumar A, Sindhu J (2019) *SAR QSAR Environ Res* 30(2):63–80. <https://doi.org/10.1080/1062936X.2018.1564067>
37. Kumar P, Kumar A, Sindhu J (2019) *SAR QSAR Environ Res* 30(8):525–541. <https://doi.org/10.1080/1062936X.2019.1629998>
38. Kumar A, Kumar P (2021) *SAR QSAR Environ Res* 32(10):817–834. <https://doi.org/10.1080/1062936X.2021.1973095>
39. Kumar P, Kumar A (2020) *SAR QSAR Environ Res* 31(9):697–715. <https://doi.org/10.1080/1062936X.2020.1806105>
40. Manisha, Chauhan S, Kumar P, Kumar A (2019) *SAR QSAR Environ Res* 30(3):145–159. <https://doi.org/10.1080/1062936X.2019.1568299>
41. Kumar A, Kumar P (2021) *Struct Chem* 32(1):149–165. <https://doi.org/10.1007/s11224-020-01629-2>
42. Nimbhal M, Bagri K, Kumar P, Kumar A (2020) *Struct Chem* 31(2):831–839. <https://doi.org/10.1007/s11224-019-01468-w>
43. Ahmadi S, Lotfi S, Kumar P (2022) *Toxicol Mech Methods* 32(4):302–312. <https://doi.org/10.1080/15376516.2021.2000686>
44. Kumar A, Bagri K, Kumar P (2020) *Drug Res* 70(5):226–232. <https://doi.org/10.1055/a-1138-8725>
45. Kumar P, Kumar A (2018) *Drug Res* 68(4):189–195. <https://doi.org/10.1055/s-0043-119288>
46. Duhan M, Kumar P, Sindhu J, Singh R, Devi M, Kumar A, Kumar R, Lal S (2021) *Comput Biol Med* 138:104876. <https://doi.org/10.1016/j.combiomed.2021.104876>
47. Kumar P, Kumar A (2020) *J Biomol Struct Dyn* 38(11):3296–3306. <https://doi.org/10.1080/07391102.2019.1656109>
48. Kumar A, Kumar P (2020) *Arch Toxicol* 94(9):3069–3086. <https://doi.org/10.1007/s00204-020-02828-w>
49. Tropsha A, Cho SJ, Zheng W (1999) In rational drug design. ACS Symposium Series, vol 719. American Chemical Society, pp 198–211. <https://doi.org/10.1021/bk-1999-0719.ch013>
50. Golbraikh A, Tropsha A (2002) *J Mol Graph Model* 20(4):269–276. [https://doi.org/10.1016/S1093-3263\(01\)00123-1](https://doi.org/10.1016/S1093-3263(01)00123-1)
51. Shayanfar A, Shayanfar S (2014) *Eur J Pharm Sci* 59:31–35. <https://doi.org/10.1016/j.ejps.2014.03.007>
52. Chirico N, Gramatica P (2011) *J Chem Inf Model* 51(9):2320–2335. <https://doi.org/10.1021/ci200211n>
53. Schüürmann G, Ebert R-U, Chen J, Wang B, Kühne R (2008) *J Chem Inf Model* 48(11):2140–2145. <https://doi.org/10.1021/ci800253u>
54. Roy K, Kar S (2014) *Eur J Pharm Sci* 62:111–114. <https://doi.org/10.1016/j.ejps.2014.05.019>
55. Lawrence IKL (1992) *Biometrics* 48(2):599–604. <https://doi.org/10.2307/2532314>

56. Consonni V, Ballabio D, Todeschini R (2010) *J Chemom* 24(3–4):194–201. <https://doi.org/10.1002/cem.1290>
57. Chai T, Draxler RR (2014) *Geosci Model Dev* 7(3):1247–1250. <https://doi.org/10.5194/gmd-7-1247-2014>
58. Kumar P, Singh R, Kumar A, Toropova AP, Toropov AA, Devi M, Lal S, Sindhu J, Singh D (2022) *SAR QSAR Environ Res* 33(9):677–700. <https://doi.org/10.1080/1062936X.2022.2120068>
59. Roy K, Mitra I, Kar S, Ojha PK, Das RN, Kabir H (2012) *J Chem Inf Model* 52(2):396–408. <https://doi.org/10.1021/ci200520g>
60. Kumar A, Kumar P, Singh D (2022) *Chemometr Intell Lab Syst* 224:104552. <https://doi.org/10.1016/j.chemolab.2022.104552>
61. Kumar P, Kumar A (2020) *Chemometr Intell Lab Syst* 200:103982. <https://doi.org/10.1016/j.chemolab.2020.103982>
62. Singh R, Kumar P, Devi M, Lal S, Kumar A, Sindhu J, Toropova AP, Toropov AA, Singh D (2022) *New J Chem* 46:19062–19072. <https://doi.org/10.1039/D2NJ03515D>
63. Kumar A, Kumar P (2021) *J Hazard Mater* 402:123777. <https://doi.org/10.1016/j.jhazmat.2020.123777>

Index

A

Absorption, Distribution, Metabolism, and Excretion (ADME), 243, 246, 247
Advantages of Quasi-SMILES, 278
AlvaDesc, 51, 121, 122, 125–128, 131–134
Amino acids, 41, 61, 87, 250, 269, 270, 274, 275, 277–285, 287, 289, 290
Anti-dengue virus activity, 117–119
Antimicrobial peptides, 271, 279, 281, 282, 288, 289
Applicability domain, 11, 60, 124, 173, 192, 193, 203, 333, 340, 387, 399, 441, 446, 455
Artificial neural network, 9, 147, 167, 172, 275, 288, 301, 303, 305, 306, 318, 374, 377
Atom pairs proportions (ATOMPAIR), 69, 276
Azo dyes, 337, 410

B

Balaban index, 32, 59
Balance of correlations, 69, 71, 77, 78, 175, 176, 277, 321, 344, 441, 444, 459
Binary liquid mixtures, 345
Bioactivity descriptors, 117, 119, 125, 131, 133, 134
Biological activity, 6, 15, 25, 27, 33, 35, 38, 69, 118, 121, 133, 182, 191, 247, 271, 272, 277, 279, 281, 282, 286, 298, 315, 332, 341, 354, 357, 361, 367, 398, 400, 409, 411, 413, 418
BOND, 65, 67, 276, 440

C

Cell viability, 47, 198, 206, 207, 317, 318, 335, 410
Chance correlations, 6, 8, 182, 305
Chemoinformatic, 25, 27, 49, 57, 119, 244, 297
Chemometrics, 25, 43, 49, 271, 378, 387
Classification model, 241, 249, 251, 252, 262, 265, 313, 318, 335
CODESSA software, 50, 180, 181, 339, 400
Comparative Molecular Field Analysis (CoMFA), 60, 247, 285
Comparative Molecular Similarity Indices Analysis (CoMSIA), 60, 285
Concordance Correlation Coefficient (CCC), 18–20, 162, 192, 204, 231, 322, 374, 380, 397, 402, 407, 408, 415–417, 446–450, 454, 458
Conductive polymer composites, 211, 212, 214, 225, 230, 235
Consensus modelling, 209, 319, 338, 344, 421–423, 442, 455, 458
CORAL software, 13, 16, 19, 43, 45, 50, 57, 62, 67, 160, 163, 191, 194–196, 200, 201, 203–206, 230, 243, 252, 281, 315, 319, 320, 322, 328, 329, 337–339, 344, 346, 351, 354, 365–367, 387, 408, 421, 422, 455, 459
Correlation coefficients, 6, 14, 17, 59, 75, 77, 144, 146, 148, 176, 202, 204, 229, 247, 251, 259, 265, 280, 281, 285, 288, 298, 330, 336, 342, 366, 374, 383, 385, 402, 404, 406–409, 414, 441, 454

- Correlation Intensity Index (CII), 75,
77–79, 159, 162, 167, 176–181, 183,
192, 202, 231, 318, 333, 335, 337,
338, 343–345, 374, 380, 384–387,
389, 391, 397, 402, 406–408,
410–412, 415–417, 421–423, 441,
444, 446–450, 454, 455, 459
- Correlation Weights (CW), 17, 50, 67, 71,
75, 146, 148, 149, 153, 160, 175,
177, 178, 183, 201, 203, 205, 209,
226, 228–230, 232–234, 241, 252,
253, 265, 276–278, 280, 282, 315,
316, 319, 322, 328, 330, 331, 333,
338, 340, 341, 343, 344, 365, 366,
374, 380, 411, 421, 422, 440, 455
- Criterion of the predictive potential, 204,
402
- Cross-validation, 59, 122, 125, 126, 244,
245, 285, 311, 319–321, 398, 399,
442
- CurlySMILES, 61
- Cytotoxicity, 14, 121, 196, 277, 297,
313–318, 321, 323, 329, 335, 409
- D**
- Daphnia magna, 46, 49, 62, 199, 206, 334
- Data collection, 49, 193, 273, 298
- Data curation, 10, 250
- Dataset, 88, 90, 99–104, 106, 108, 117,
119–122, 125, 133, 147, 148, 162,
193, 195, 196, 200, 205, 241, 243,
244, 250, 251, 265, 273, 274, 277,
279, 287, 289, 290, 318, 320–322,
335, 340, 341, 343, 344, 385,
387–391, 399, 421, 423
- 0D descriptors, 25, 30, 48, 329
- 1D descriptors, 7, 10, 25, 30, 48
- 2D descriptors, 25, 30, 48, 181, 245, 280,
328, 329, 338, 357
- 3D descriptors, 8, 25, 33, 34, 48, 280, 357
- 4D descriptors, 8, 25, 34, 48
- 7D descriptors, 329
- Deep neural networks, 61
- DeepSMILES, 61
- Descriptors for nano-QSPR/QSAR, 42
- Disadvantages of Quasi-SMILES, 278
- DRAGON descriptors, 50, 180, 181, 400
- Drug design, 4, 88, 192, 327, 346
- DTC-QSAR, 401
- E**
- Eclectic information, 163, 191, 269, 275
- Electrophilicity index, 144
- Empirical & experimental descriptors, 39
- Enalos InSilicoNano platform, 354
- Energy, 12, 32, 35–37, 39, 41, 42, 90, 141,
143, 145, 169, 172, 214, 222, 224,
225, 318, 327, 328, 343, 376, 377,
398
- Epitope peptides, 279, 282, 283
- Evidence Lower Bound (ELBO), 89,
94–96, 101, 106
- F**
- Fisher's statistics, 123
- Flammability properties, 297, 318
- Food property prediction, 340
- Fullerenes C60 and C70, 357, 358
- Fuzzy set, 182, 404
- G**
- Gated Recurrent Units (GRUs), 89, 91–93,
96, 107
- Genetic Algorithm (GA), 126, 181, 274,
284–287, 290, 301, 303, 401
- Genetic algorithm-based multiple-linear
regression, 122
- Genotoxicity, 247, 334, 362
- Geometry, Topology, Atom-Weights
Assembly, 34
- Gibbs Free Energy, 343
- Glass transition temperature, 167, 169, 171,
178–180, 343, 344
- Global graph invariants, 67
- Global SMILES attributes, 65, 177
- Graph of atomic orbitals (GAO), 15–17, 67,
74, 146, 148, 159, 160, 163
- H**
- HALO, 65, 69, 276, 440
- Hammett constants, 143, 144, 147
- HBACE-1 inhibitors, 345
- Higher order connectivity, 31
- Human intestinal transporter, 241, 243,
244, 248, 249, 262, 263, 265
- Hybrid optimal descriptors, 15, 16, 159,
163, 167, 177, 179, 319–321, 340
- Hydrogen-Filled Graph (HFG), 67, 321
- Hydrogen Suppressed Graph (HSG), 15, 16,
58, 67, 73, 319–321, 328, 337, 343

I

Index of Ideality of Correlation (IIC),
17–20, 75, 78, 79, 148, 159, 162,
167, 176, 177, 179–181, 183, 192,
202, 231, 241, 251, 252, 263, 265,
283, 319, 321, 322, 328, 329,
332–346, 374, 380, 384, 386, 387,
397, 402, 404–412, 415–417,
446–450, 454
In silico, 87, 242–244, 246–250, 271, 272,
290, 328, 340
International Chemical Identifier (InChI),
57, 58, 60
In vitro, 33, 242, 245–250, 315
In vivo, 242, 246–248, 315
Ionic liquids, 297, 319–323, 376, 379
ISA-TAB-Nano paradigm, 353

K

K-Nearest Neighbour classification, 9

L

Latent space optimization, 96
Long Short-Term Memories (LSTMs), 89,
91, 94–96, 101

M

Mathematical modelling, 4, 5, 271, 284
Matthews correlation coefficient (MCC),
241, 251, 252, 259, 262–265, 281
Mechanistic interpretation, 7, 118, 126,
173, 193, 205, 313, 316, 320, 321,
333, 341, 343, 345, 455, 459
Micelle-polymer, 341
Model evaluation, 298
Model validation, 11, 142, 200, 204, 273,
322
Modified Hosoya index (Z^*), 32
Molecular descriptors, 3, 6, 7, 16, 25–30,
32, 34, 35, 39, 46, 48–50, 59, 60,
119, 121, 125, 142, 143, 145, 148,
163, 171, 172, 298–300, 303, 304,
320, 323, 340, 400, 422
Molecular graph, 6, 13, 31, 43, 57–59, 67,
69, 85–93, 96, 99, 112, 139, 144,
145, 177, 179, 180, 193, 195, 228,
329, 340, 358
Molecular optimization, 88, 94, 96, 98, 99,
101, 104, 108
Molecular orbital energies, 36

Molecular structure, 3, 5–7, 13, 26, 27, 29,
30, 34, 36, 43, 57–61, 63, 65, 67, 69,
121, 142, 145, 160, 168, 172, 174,
177, 182, 183, 191, 193, 195, 226,
228, 244, 250, 251, 269, 272, 273,
275, 276, 279, 281, 297–299, 301,
311, 313, 318, 319, 321, 329, 330,
332, 334, 335, 338, 340, 341, 343,
345, 352, 354, 356–358, 373, 378,
379, 398, 400, 401
Monte Carlo Method, 10, 12, 14, 15, 69,
71, 146, 148, 178, 230, 278, 281,
315, 316, 327, 330, 331, 333, 334,
337, 339, 343, 346, 365, 366, 380,
384, 386, 390, 409, 411
Mordred, 121, 125–127
Multi-Layer Perception (MLP), 131, 374,
378
Multiple Linear Regressions (MLR), 8,
126, 167, 171, 180, 274, 284,
299–301, 303–306, 356, 401
Multiwalled carbon nanotubes, 207, 208,
329
Mutagenicity, 228, 329, 332, 333, 352, 358

N

Nanofluids, 207, 208, 334, 335, 373,
375–392
Nanomaterials, 42–44, 49, 61, 62, 182, 191,
205, 222, 313–317, 322, 323, 327,
328, 332, 334, 346, 351–359, 361,
363, 366, 367
Nanoparticles, 12, 14, 42, 43, 46, 49, 62,
169, 192, 196, 197, 213, 222, 297,
313–318, 323, 328, 329, 332–336,
346, 354, 359, 373–379, 382–386,
388, 389, 400, 411, 418
Nano-QSPR/QSAR, 42, 43, 49, 61, 62,
193, 205, 332, 366
Nanotechnology, 42, 169, 192, 314, 315,
318, 335, 352–355, 376
Non-linear regression, 124, 173, 309
NOSP, 65, 71, 276, 440
Number of iterations or epochs, 71, 123,
160

O

Octanol-water partition coefficient, 42, 60,
143, 299, 335, 336, 418
OECD principles, 3, 6, 7, 173, 205, 332,
442

Optimal descriptor, 13, 14, 17–20, 44, 45, 57, 62–64, 67, 71, 139, 145–147, 149, 153, 163, 173, 175–178, 226, 252, 260, 277, 278, 316, 317, 322, 328, 329, 334, 338, 339, 341, 343, 365, 374, 380, 386, 387, 406, 407, 440

P

PaDEL descriptors, 181
Partial Least Squares (PLS), 8, 9, 147, 167, 172, 180, 181, 274, 285, 290, 301, 303–306, 318, 398, 401
Particle Swarm Optimization Algorithm (PSO), 287, 289
Peptides, 41, 61, 100, 242, 245, 247, 251, 269–282, 284–291, 329, 361, 367
Percolation threshold, 211, 212, 214, 215, 225–228, 230, 231, 235
Pharmacodynamics, 275
Pharmacokinetics, 30, 241, 242, 247, 248, 250, 275
Polarizability ZZ index, 145
Polymer, 106, 167–173, 178, 179, 182, 183, 211–214, 216–227, 229, 230, 235, 236, 329, 341–343, 354, 418
Prediction model, 242, 246, 249, 298, 313, 315, 316, 318, 323, 338, 341, 342
Principal component analysis (PCA), 8, 9, 41, 245, 274, 285, 304
Property prediction, 88, 94, 96, 99, 100, 102–104, 108, 112
Protease inhibitor, 15, 118, 119
PyDescriptors, 121, 122, 125–127

Q

QSARINS, 181, 285
Quantum chemical descriptors, 35, 36, 39
Quasi-SMILES descriptors, 43, 45, 50, 193, 196, 275, 297, 311, 317, 318, 335, 342, 390

R

Random Forest, 10, 100, 167, 244, 249, 382, 398, 401
Random processes, 69
Rdkit, 50, 96, 98, 121, 125–128, 132–134
Recommendations for building robust QSPR/QSAR models, 12
Recurrent Neural Networks (RNNs), 9, 85–88, 90–92, 96, 110–112

Redox potential, 139, 141–147, 163
Refractive index, 167, 169, 178–180, 320
Reproducibility, 4, 11, 12, 19, 21, 69, 182
Retention index, 344, 422, 423, 457
Risk assessment, 42, 317, 327, 332, 333, 335–337, 346, 353, 356

S

Self-accelerating decomposition temperature, 328, 341
Self-consistent model system, 336, 339, 346, 397, 414, 418
SFS-QSAR-tool, 122
Siamese Neural Network (SNN), 119, 120
Signaturizer, 117, 119, 121, 125–128, 131–134
SIRMS, 121, 125–128, 132–134
Skin sensitivity, 411
SMiles Arbitrary Target Specification (SMARTS), 57, 58, 60
SMILES descriptors, 43, 57, 275, 313, 321, 322
SmilesDrawer, 61
SMILES strings, 61–65, 85–92, 94–96, 98, 100, 101, 106–112, 195, 277, 312, 316, 366, 379
Software for generation of molecular descriptors, 46
Statistical characteristics, 14, 16, 18, 162, 177, 182, 230, 231, 321, 366, 385, 386, 401, 409, 411, 412
Stepwise Multiple Regression-Partial Least Square regression (SMR-PLS), 285
Support Vector Machine (SVM), 9, 10, 100, 131, 168, 173, 180, 181, 244, 275, 287, 289, 290, 298, 303, 307–311, 313, 318, 354, 374, 377

T

Target function, 17, 75, 148, 178, 192, 201, 251, 277, 282, 283, 317, 332–338, 342–344, 366, 374, 383, 385, 387, 391, 411, 421, 441, 444, 446, 459
Thermophysical properties, 373, 376, 377, 379, 380, 382, 385–387, 392
Threshold, 17, 71, 159, 175, 203, 211, 229, 233–235, 276, 284, 366, 440, 444, 455
Topological indices, 31, 32, 59, 145, 355
Topological Maximum Cross Correlation (TMACC), 32

Toxicity, [4–6](#), [15](#), [16](#), [20](#), [27](#), [33](#), [35](#), [42](#), [46](#),
[63](#), [78](#), [97](#), [100](#), [102](#), [103](#), [107](#), [168](#),
[192](#), [196](#), [199](#), [228](#), [247](#), [297](#), [298](#),
[312–318](#), [321–323](#), [328](#), [329](#), [331](#),
[333–335](#), [337](#), [338](#), [346](#), [354](#), [356](#),
[411](#)

Types of molecular descriptors, [29](#), [143](#),
[299](#)

V

Validity, [100](#), [101](#), [147](#), [163](#), [205](#), [345](#), [380](#),
[402](#)

Variational autoencoders, [61](#), [85](#), [86](#), [88](#),
[89](#), [101](#), [102](#)

VEGA, [400](#)

Virtual Computational Chemistry
Laboratory, [181](#), [400](#)

W

Weighted Hlistic invariant molecular, [34](#)

Wiener (W) index, [30](#), [31](#)

Williams plots, [124](#), [128](#), [130](#)

Z

Zagreb indices, [32](#)