

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**

|   |     |
|---|-----|
| <b>16 On Complementary Approaches of Assessing the Predictive Potential of QSPR/QSAR Models</b> .....   | 397 |
| Andrey A. Toropov, Alla P. Toropova, Danuta Leszczynska,<br>and Jerzy Leszczynski   |     |
| <b>17 CORAL: Predictions of Quality of Rice Based on Retention Index Using a Combination of Correlation Intensity Index and Consensus Modelling</b> ..... | 421 |
| Parvin Kumar and Ashwani Kumar  |     |
| <b>Index</b> .....  | 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                    |
| $\Delta G^\ddagger$ | Gibb's activation free energy         |

## Greek Symbols

|           |                                     |
|-----------|-------------------------------------|
| $\rho$    | Density                             |
| $\varphi$ | 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     |

|                         |                          |
|-------------------------|--------------------------|
| $\text{In}_2\text{O}_3$ | Indium (III) oxide       |
| $\text{La}_2\text{O}_3$ | Lanthanum oxide          |
| $\text{MgO}$            | Magnesium oxide          |
| $\text{Mn}_2\text{O}_3$ | Manganese (III) oxide    |
| $\text{Mn}_3\text{O}_4$ | Manganese (II,III) oxide |
| $\text{Ni}_2\text{O}_3$ | Nickel (III) oxide       |
| $\text{NiO}$            | Nickel (II) oxide        |
| $\text{Sb}_2\text{O}_3$ | Antimony oxide           |
| $\text{Si}_3\text{N}_4$ | Silicon nitride          |
| $\text{SiC}$            | Silicon carbide          |
| $\text{SiO}_2$          | Silicon dioxide          |
| $\text{SnO}_2$          | Tin (IV) oxide           |
| $\text{TiN}$            | Titanium nitride         |
| $\text{TiO}_2$          | Titanium dioxide         |
| $\text{WO}_3$           | Tungsten (VI) oxide      |
| $\text{Y}_2\text{O}_3$  | Yttrium (III) oxide      |
| $\text{Yb}_2\text{O}_3$ | Ytterbium (III) oxide    |
| $\text{ZnO}$            | Zinc oxide               |
| $\text{ZrO}_2$          | Zirconium oxide          |



# 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](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](mailto:ncordeir@fc.up.pt)