

Contents

Preface *XIII*

List of Contributors *XV*

1	Current Modeling Methods Used in QSAR/QSPR	1
	<i>Liew Chin Yee and Yap Chun Wei</i>	
1.1	Introduction	1
1.2	Modeling Methods	3
1.2.1	Methods for Regression Problems	3
1.2.1.1	Multiple Linear Regression	3
1.2.1.2	Partial Least Squares	4
1.2.1.3	Feedforward Backpropagation Neural Network	5
1.2.1.4	General Regression Neural Network	7
1.2.1.5	Gaussian Processes	9
1.2.2	Methods for Classification Problems	10
1.2.2.1	Logistic Regression	10
1.2.2.2	Linear Discriminant Analysis	11
1.2.2.3	Decision Tree and Random Forest	12
1.2.2.4	<i>k</i> -Nearest Neighbor	14
1.2.2.5	Probabilistic Neural Network	15
1.2.2.6	Support Vector Machine	16
1.3	Software for QSAR Development	18
1.3.1	Structure Drawing or File Conversion	19
1.3.2	3D Structure Generation	19
1.3.3	Descriptor Calculation	20
1.3.4	Modeling	21
1.3.5	General purpose	23
1.4	Conclusion	24
	References	26

2	Developing Best Practices for Descriptor-Based Property Prediction: Appropriate Matching of Datasets, Descriptors, Methods, and Expectations	33
	<i>Michael Krein, Tao-Wei Huang, Lisa Morkowchuk, Dimitris K. Agrafiotis, and Curt M. Breneman</i>	
2.1	Introduction	33
2.1.1	Posing the Question	34
2.1.2	Validating the Models	35
2.1.3	Interpreting the Models	36
2.2	Leveraging Experimental Data and Understanding their Limitations	36
2.3	Descriptors: The Lexicon of QSARs	37
2.3.1	Classical QSAR Descriptors and Uses	38
2.3.2	Experimentally Derived Descriptors	38
2.3.2.1	Biodescriptors	39
2.3.2.2	Descriptors from Spectroscopy/Spectrometry and Microscopy	40
2.3.3	0D, 1D and 2D Computational Descriptors	40
2.3.4	3D Descriptors and Beyond	41
2.3.5	Local Molecular Surface Property Descriptors	42
2.3.6	Quantum Chemical Descriptors	42
2.4	Machine Learning Methods: The Grammar of QSARs	44
2.4.1	Principal Component Analysis	44
2.4.2	Factor Analysis	45
2.4.3	Multidimensional Scaling, Stochastic Proximity Embedding, and Other Nonlinear Dimensionality Reduction Methods	45
2.4.4	Clustering	46
2.4.5	Partial Least Squares (PLS)	47
2.4.6	<i>k</i> -Nearest Neighbors (kNN)	47
2.4.7	Neural Networks	48
2.4.8	Ensemble Models	49
2.4.9	Decision Trees and Random Forests	49
2.4.10	Kernel Methods	50
2.4.11	Ranking Methods	52
2.5	Defining Modeling Strategies: Putting It All Together	52
2.6	Conclusions	56
	References	57
3	Mold² Molecular Descriptors for QSAR	65
	<i>Huixiao Hong, Svetoslav Slavov, Weigong Ge, Feng Qian, Zhenqiang Su, Hong Fang, Yiyu Cheng, Roger Perkins, Leming Shi, and Weida Tong</i>	
3.1	Background	65
3.1.1	History of QSAR	65
3.1.2	Introduction to QSAR	67
3.1.3	Molecular Descriptors: Bridge for QSAR	68
3.1.3.1	Molecular Descriptors	69
3.1.3.2	Role of Molecular Descriptors	70

3.1.3.3	Types of Molecular Descriptors	71
3.1.3.4	Calculation of Molecular Descriptors (Software Packages)	71
3.2	Mold ² Molecular Descriptors	71
3.2.1	Description of Mold ² Descriptors	73
3.2.1.1	Topological Descriptors	73
3.2.1.2	Constitutional Descriptors	94
3.2.1.3	Information Content-based Descriptors	94
3.2.2	Calculation of Mold ² Descriptors	94
3.2.3	Evaluation of Mold ² Descriptors	96
3.2.3.1	Information Content by Shannon Entropy Analysis	96
3.2.3.2	Correlations between Descriptors	98
3.3	QSAR Using Mold ² Descriptors	99
3.3.1	Classification Models based on Mold ² Descriptors	100
3.3.2	Regression Models based on Mold ² Descriptors	102
3.4	Conclusion Remarks	105
	References	105
4	Multivariate Analysis of Molecular Descriptors	111
	<i>Viviana Consonni and Roberto Todeschini</i>	
4.1	Introduction	111
4.2	2D Matrix-Based Descriptors	114
4.3	Graph-Theoretical Matrices	120
4.3.1	Vertex Weighting Schemes	122
4.4	Multivariate Similarity Analysis of Chemical Spaces	122
4.5	Analysis of Chemical Information of Descriptors from Graph-Theoretical Matrices	124
4.5.1	Data Sets	124
4.5.2	Comparison of Graph-Theoretical Matrices	125
4.5.2.1	Comparison of Weighted Graph-Theoretical Matrices	130
4.5.3	Comparison of Matrix Operators	133
4.5.4	Comparison of Single Operators from Different Graph-Theoretical Matrices	137
4.6	Conclusions	143
	References	143
5	Partial-Order Ranking and Linear Modeling: Their Use in Predictive QSAR/QSPR Studies	149
	<i>Andrew G. Mercader and Eduardo A. Castro</i>	
5.1	Introduction	149
5.2	Linear QSAR Methodology, ERM, RM and GA	150
5.2.1	Replacement Method	153
5.2.2	Enhanced Replacement Method	154
5.2.3	Genetic Algorithm	154
5.2.4	Main Differences between MRM and RM	156
5.3	Principles of Ranking Methods	159

5.4	Selection of the Molecular Descriptors for Ranking	163
5.5	QSAR Based on Hasse Diagrams	165
5.6	Discussion	165
5.7	Conclusions	169
	References	170
6	Graph-Theoretical Descriptors for Branched Polymers	175
	<i>Koh-Hei Nitta</i>	
6.1	Introduction	175
6.2	Algebraic Graph Theory	176
6.3	Ideal Chain Models	180
6.4	Graph-Theoretical Approach to Chain Dynamics and Statistics	182
6.4.1	Radius of Gyration	182
6.4.2	Rouse Dynamics	185
6.4.3	Intrinsic Viscosity	188
6.4.4	Scattering Function	190
6.4.5	High Moments of Relaxation Time and Radius of Gyration	191
6.5	Applications	193
6.6	Final Remarks	194
	References	196
7	Structural-Similarity-Based Approaches for the Development of Clustering and QSPR/QSAR Models in Chemical Databases	201
	<i>Irene Luque Ruiz, Gonzalo Cerruela García, and Miguel Ángel Gómez-Nieto</i>	
7.1	Chemical Structural Similarity	201
7.1.1	Molecular Graph and Structural Similarity	203
7.1.2	Descriptor-Based Structural Similarity	203
7.1.3	Combining Structural Similarity Approaches	204
7.1.4	Approximate Structural Similarity	205
7.2	Clustering Models Based on Structural Similarity	207
7.2.1	Clustering of Chemical Databases	211
7.2.1.1	Pattern Representation of Chemicals Structures	211
7.2.1.2	Clustering of Chemical Databases	212
7.3	QSPR/QSAR Models Based on Structural Similarity	217
7.3.1	Dataset Selection	219
7.3.2	Dataset Representation	220
7.3.3	Fitting of the Dataset Representation	221
7.3.4	Building and Validation of the QSAR Model	221
	References	223
8	Statistical Methods for Predicting Compound Recovery Rates for Ligand-Based Virtual Screening and Assessing the Probability of Activity	229
	<i>Martin Vogt and Jürgen Bajorath</i>	
8.1	Introduction	229
8.2	Theory	231

8.2.1	Bayesian Approach to Virtual Screening	231
8.2.2	Predicting the Performance of Bayesian Screening	235
8.2.3	Practical Prediction of Compound Recall	236
8.2.4	Exemplary Results	238
8.3	Alternative Approaches to the Prediction of Compound Recall	238
8.4	Conclusions	240
	References	241
9	Molecular Descriptors and the Electronic Structure	245
	<i>Bögel Horst</i>	
9.1	Introduction	245
9.2	The Structure of Molecules	246
9.2.1	General Remarks	246
9.2.2	Structure Coding	247
9.2.3	Structural Features	248
9.2.4	Structure and Energy	250
9.3	The Electronic Structure	251
9.4	Dividing Molecules in Atoms and Bonds	254
9.4.1	Bonding in Molecules	254
9.4.2	Energy Partitioning	255
9.4.3	Energy and the Hückel Approach	255
9.4.4	Energy Components of Atoms and Bonds	256
9.4.5	Perturbation Treatment of the Electronic Structure	257
9.4.6	Thermodynamic Equilibrium	258
9.4.7	Model of "Atom in Molecules"	258
9.5	Structure and Dynamics	259
9.5.1	Molecular Flexibility	259
9.5.2	Molecular Dynamics Simulation	259
9.5.3	Conformational Space	260
9.6	Structure and Properties	262
9.6.1	Structure Property Relationships	262
9.6.2	Type of Molecular Properties	262
9.6.3	Molecular Commonality and Similarity	263
9.6.4	Multilinear Regression	263
9.6.5	Selection of Molecular Descriptors	265
9.7	Modeling of Physicochemical Properties of the Isomers of Hexane	265
9.8	Modeling of the Proton Affinity	275
9.8.1	Proton Affinity of Pyridines	275
9.8.1.1	Data and Mechanism	275
9.8.1.2	Model I	277
9.8.1.3	Model II	278
9.8.1.4	Model III	280
9.8.1.5	Model IV	281
9.8.1.6	Model V	281
9.8.1.7	Model VI	282

9.8.2	Basicity of N-Heterocyclic Aromatics	283
9.9	Molecular Surface Properties	285
9.10	Conclusions	290
	References	291
10	New Types of Descriptors and Models in QSAR/QSPR	293
	<i>Christian Kramer and Timothy Clark</i>	
10.1	Introduction	293
10.2	Local Properties	294
10.2.1	Molecular Electrostatic Potential	294
10.2.2	Electron Density	295
10.2.3	Local Polarizability	295
10.2.4	Local Ionization Energy and Local Electron Affinity	296
10.3	Descriptors Derived from Local Properties	297
10.3.1	PEST Methodology	297
10.4	MEP as Descriptor for Hydrogen-Bonding Strengths	298
10.5	ParaSurf (Politzer–Murray) Descriptors	298
10.6	4D: Conformational-Ensemble-based Descriptors	299
10.7	Proper Validation/Generation of QSA(P)R Models	300
10.8	Conclusions	302
	References	303
11	Consensus Models of Activity Landscapes	307
	<i>José L. Medina-Franco, Austin B. Yongye, and Fabian López-Vallejo</i>	
11.1	Introduction	307
11.2	Characterization of the Activity Landscape	309
11.3	Consensus Models of Activity Landscape	312
11.3.1	Chemical Space and Molecular Representation	312
11.3.2	Activity Landscape with Multiple Representations	316
11.4	Conclusions and Future Perspectives	322
	References	323
12	Reverse Engineering Chemical Reaction Networks from Time Series Data	327
	<i>Dominic P. Searson, Mark J. Willis, and Allen Wright</i>	
12.1	Introduction	327
12.2	Problem Definition	329
12.3	Reconstruction of Elementary Reaction Networks from Data by Network Search	331
12.3.1	Network Search as a Nonlinear Integer Programming Problem	332
12.3.2	Estimation of the Rate Coefficients for Trial Reaction Networks	333
12.4	Formulation of the Objective Function for Network Search	335
12.4.1	Physical/Chemical Information Available	336
12.4.2	No physical/Chemical Information Available	336

12.5	Differential Evolution for Searching the Space of Reaction Networks	337
12.5.1	Basic DE Optimization Method	338
12.5.2	Self-Adaptive DE with Integer Variables	339
12.6	Network Identification Case Studies	340
12.6.1	Estimation of Time Derivatives	342
12.6.2	DE Settings	343
12.6.3	Model Selection Methodology	343
12.6.4	Results	344
12.7	Conclusions	346
	References	347
13	Reduction of Dimensionality, Order, and Classification in Spaces of Theoretical Descriptions of Molecules: An Approach Based on Metrics, Pattern Recognition Techniques, and Graph Theoretic Considerations	349
	<i>George Maroulis</i>	
13.1	Introduction	349
13.2	Theory	351
13.3	Methods and Computational Strategy	354
13.4	Results and Discussion	358
13.5	Conclusions	363
	References	363
14	The Analysis of Organic Reaction Pathways by Brownian Processing	365
	<i>Daniel J. Graham</i>	
14.1	Introduction	365
14.2	Electronic Messages, Information, and Energy	366
14.3	Molecular Messages, Conversions, and State Space Representations	374
14.4	Closing	389
	References	390
15	Generation of Chemical Transformations: Reaction Pathways Prediction and Synthesis Design	393
	<i>Grażyna Nowak and Grzegorz Fic</i>	
15.1	Introduction	393
15.2	The Graph Transformation Rules for Generation of Chemical Reactions	396
15.2.1	The Graph-Theoretic Reaction Rules and Formal-Logical Approach for Reaction Generation	397
15.2.1.1	The Chemical Reaction Graph	399
15.2.1.2	Ugi and Dugundji Formal Theory for Reactions and Reaction Mechanisms	400

15.2.2	The Empirical Reaction Rules and Knowledge-Based Approach for Reaction Generation. Automated Creation of Rules by Learning and Reaction Database Mining	404
15.2.2.1	Automatically Derived Reaction Rules	404
15.2.2.2	Functional Group Transformations	406
15.2.2.3	Substructure-Based Transformations	406
15.3	Combinatorial Complexity Problem: Strategies for the Directed Reaction Generation	409
15.3.1	Retrosynthetic Generation of Chemical Transformations: Computer-Assisted Synthesis Design	410
15.3.1.1	Recognition of Guiding Patterns, Molecular Symmetry, or Isomorphic Substructures	411
15.3.1.2	Complexity-Based Disconnective Strategies	412
15.3.1.3	Concept of the Strategic Bond Tree for Disconnections	413
15.3.2	Forward Generation of Chemical Transformations: Computer-Assisted Reaction Prediction	414
15.3.2.1	Quantitative Models for Reactivity Prediction	416
15.3.2.2	Formal-Logical Approach to the Search Space of Possible Chemical Transformations	418
15.4	Conclusion	419
	References	420
	Index	427