Contents

Foreword *xvii*
List of Contributors *xxi*

1 Introduction *1*
Thomas Engel and Johann Gasteiger
1.1 The Rationale for the Books *1*
1.2 Development of the Field *2*
1.3 The Basis of Chemoinformatics and the Diversity of Applications *3*
1.3.1 Databases *3*
1.3.2 Fundamental Questions of a Chemist *4*
1.3.3 Drug Discovery *5*
1.3.4 Additional Fields of Application *6*
Reference *7*

2 QSAR/QSPR *9*
Wolfgang Sippl and Dina Robaa
2.1 Introduction *9*
2.2 Data Handling and Curation *13*
2.2.1 Structural Data *13*
2.2.2 Biological Data *14*
2.3 Molecular Descriptors *14*
2.3.1 Structural Keys (1D) *15*
2.3.2 Topological Descriptors (2D) *16*
2.3.3 Geometric Descriptors (3D) *16*
2.4 Methods for Data Analysis *17*
2.4.1 Overview *17*
2.4.2 Unsupervised Learning *17*
2.4.3 Supervised Learning *18*
2.5 Classification Methods *19*
2.5.1 Principal Component Analysis *19*
2.5.2 Linear Discriminant Analysis *19*
2.5.3 Kohonen Neural Network *19*
2.5.4 Other Classification Methods *20*
2.6 Methods for Data Modeling *20*

2.6.1 Regression-Based QSAR Approaches *20*
2.6.2 3D QSAR *22*
2.6.3 Nonlinear Models *25*
2.7 Summary on Data Analysis Methods *30*
2.8 Model Validation *30*
2.8.1 Proper Use of Validation Routines *31*
2.8.2 Modeling/Validation Workflow *32*
2.8.3 Splitting of Datasets *32*
2.8.4 Compilation of Modeling, Training, Validation, Test, and External Sets *34*
2.8.5 Cross-Validation *36*
2.8.6 Bootstrapping *37*
2.8.7 Y-Randomization (Y-Scrambling) *38*
2.8.8 Goodness of Prediction and Quality Criteria *39*
2.8.9 Applicability Domain and Model Acceptability Criteria *41*
2.8.10 Scope of External and Internal Validation *43*
2.8.11 Validation of Classification Models *45*
2.9 Regulatory Use of QSARs *46*
Selected Reading *48*
References *49*

3 Prediction of Physicochemical Properties of Compounds *53*
Igor V. Tetko, Aixia Yan, and Johann Gasteiger
3.1 Introduction *53*
3.2 Overview of Modeling Approaches to Predict Physicochemical Properties *54*
3.2.1 Prediction of Properties Based on Other Properties *55*
3.2.2 Prediction of Properties Based on Theoretical Calculations *55*
3.2.3 Additivity Schemes for Property Prediction *56*
3.2.4 Statistical Quantitative Structure–Property Relationships (QSPRs) *59*
3.3 Methods for the Prediction of Individual Properties *59*
3.3.1 Mean Molecular Polarizability *59*
3.3.2 Thermodynamic Properties *60*
3.3.3 Octanol/Water Partition Coefficient (Log *P*) *63*
3.3.4 Octanol/Water Distribution Coefficient (log *D*) *67*
3.3.5 Estimation of Water Solubility (log *S*) *69*
3.3.6 Melting Point (MP) *71*
3.3.7 Acid Ionization Constants *73*
3.4 Limitations of Statistical Methods *76*
3.5 Outlook and Perspectives *76*
Selected Reading *78*
References *78*

4 Chemical Reactions *83*

4.1 Chemical Reactions – An Introduction *84*
Johann Gasteiger
References *85*

4.2 Reaction Prediction and Synthesis Design *86*
Jonathan M. Goodman
4.2.1 Introduction *86*
4.2.2 Reaction Prediction *87*
4.2.3 Synthesis Design *94*
4.2.4 Conclusion *102*
References *103*

4.3 Explorations into Biochemical Pathways *106*
Oliver Sacher and Johann Gasteiger
4.3.1 Introduction *106*
4.3.2 The BioPath.Database *110*
4.3.3 BioPath.Explore *111*
4.3.4 Search Results *112*
4.3.5 Exploitation of the Information in BioPath.Database *117*
4.3.6 Summary *129*
Selected Reading *130*
References *130*

5 Structure–Spectrum Correlations and Computer-Assisted Structure Elucidation *133*
Joao Aires de Sousa
5.1 Introduction *133*
5.2 Molecular Descriptors *135*
5.2.1 Fragment-Based Descriptors *135*
5.2.2 Topological Structure Codes *135*
5.2.3 Three-Dimensional Molecular Descriptors *137*
5.3 Infrared Spectra *137*
5.3.1 Overview *137*
5.3.2 Infrared Spectra Simulation *138*
5.4 NMR Spectra *140*
5.4.1 Quantum Chemistry Prediction of NMR Properties *142*
5.4.2 NMR Spectra Prediction by Database Searching *142*
5.4.3 NMR Spectra Prediction by Increment-Based Methods *143*
5.4.4 NMR Spectra Prediction by Machine Learning Methods *144*
5.5 Mass Spectra *150*
5.5.1 Identification of Structures and Interpretation of MS *150*
5.5.2 Prediction of MS *151*
5.5.3 Metabolomics and Natural Products *151*
5.6 Computer-Aided Structure Elucidation (CASE) *153*

Selected Reading *157*
Acknowledgement *157*
References *158*

6.1 Drug Discovery: An Overview *165*
Lothar Terfloth, Simon Spycher, and Johann Gasteiger
6.1.1 Introduction *165*
6.1.2 Definitions of Some Terms Used in Drug Design *167*
6.1.3 The Drug Discovery Process *167*
6.1.4 Bio- and Chemoinformatics Tools for Drug Design *168*
6.1.5 Structure-based and Ligand-Based Drug Design *168*
6.1.6 Target Identification and Validation *169*
6.1.7 Lead Finding *171*
6.1.8 Lead Optimization *182*
6.1.9 Preclinical and Clinical Trials *188*
6.1.10 Outlook: Future Perspectives *189*
Selected Reading *191*
References *191*

6.2 Bridging Information on Drugs, Targets, and Diseases *195*
Andreas Steffen and Bertram Weiss
6.2.1 Introduction *195*
6.2.2 Existing Data Sources *196*
6.2.3 Drug Discovery Use Cases in Computational Life Sciences *196*
6.2.4 Discussion and Outlook *201*
Selected Reading *202*
References *202*

6.3 Chemoinformatics in Natural Product Research *207*
Teresa Kaserer, Daniela Schuster, and Judith M. Rollinger
6.3.1 Introduction *207*
6.3.2 Potential and Challenges *208*
6.3.3 Access to Software and Data *211*
6.3.4 *In Silico* Driven Pharmacognosy-Hyphenated Strategies *219*
6.3.5 Opportunities *220*
6.3.6 Miscellaneous Applications *228*
6.3.7 Limits *228*
6.3.8 Conclusion and Outlook *229*
Selected Reading *231*
References *231*

6.4 Chemoinformatics of Chinese Herbal Medicines *237*
Jun Xu
6.4.1 Introduction *237*
6.4.2 Type 2 Diabetes: The Western Approach *237*
6.4.3 Type 2 Diabetes: The Chinese Herbal Medicines Approach *238*

6.4.4 Building a Bridge *238*
6.4.5 Screening Approach *240*
Selected Reading *244*
References *244*

6.5 PubChem *245*
Wolf-D. Ihlenfeldt
6.5.1 Introduction *245*
6.5.2 Objectives *246*
6.5.3 Architecture *246*
6.5.4 Data Sources *247*
6.5.5 Submission Processing and Structure Representation *248*
6.5.6 Data Augmentation *249*
6.5.7 Preparation for Database Storage *249*
6.5.8 Query Data Preparation and Structure Searching *250*
6.5.9 Structure Query Input *253*
6.5.10 Query Processing *254*
6.5.11 Getting Started with PubChem *254*
6.5.12 Web Services *255*
6.5.13 Conclusion *255*
References *256*

6.6 Pharmacophore Perception and Applications *259*
Thomas Seidel, Gerhard Wolber, and Manuela S. Murgueitio
6.6.1 Introduction *259*
6.6.2 Historical Development of the Modern Pharmacophore Concept *260*
6.6.3 Representation of Pharmacophores *262*
6.6.4 Pharmacophore Modeling *268*
6.6.5 Application of Pharmacophores in Drug Design *272*
6.6.6 Software for Computer-Aided Pharmacophore Modeling and Screening *278*
6.6.7 Summary *278*
Selected Reading *279*
References *280*

6.7 Prediction, Analysis, and Comparison of Active Sites *283*
Andrea Volkamer, Mathias M. von Behren, Stefan Bietz, and Matthias Rarey
6.7.1 Introduction *283*
6.7.2 Active Site Prediction Algorithms *284*
6.7.3 Target Prioritization: Druggability Prediction *292*
6.7.4 Search for Sequentially Homologous Pockets *296*
6.7.5 Target Comparison: Virtual Active Site Screening *298*
6.7.6 Summary and Outlook *304*
Selected Reading *306*
References *306*

6.8 Structure-Based Virtual Screening *313*
Adrian Kolodzik, Nadine Schneider, and Matthias Rarey
6.8.1 Introduction *313*
6.8.2 Docking Algorithms *315*
6.8.3 Scoring *317*
6.8.4 Structure-Based Virtual Screening Workflow *321*
6.8.5 Protein-Based Pharmacophoric Filters *323*
6.8.6 Validation *323*
6.8.7 Summary and Outlook *326*
Selected Reading *328*
References *328*

6.9 Prediction of ADME Properties *333*
Aixia Yan
6.9.1 Introduction *333*
6.9.2 General Consideration on SPR/QSPR Models *334*
6.9.3 Estimation of Aqueous Solubility (log *S*) *336*
6.9.4 Estimation of Blood–Brain Barrier Permeability (log *BB*) *342*
6.9.5 Estimation of Human Intestinal Absorption (HIA) *346*
6.9.6 Other ADME Properties *349*
6.9.7 Summary *354*
Selected Reading *355*
References *355*

6.10 Prediction of Xenobiotic Metabolism *359*
Anthony Long and Ernest Murray
6.10.1 Introduction: The Importance of Xenobiotic Biotransformation in the Life Sciences *359*
6.10.2 Biotransformation Types *362*
6.10.3 Brief Review of Methods *364*
6.10.4 User Needs: Scientists Use Metabolism Information in Different Ways *370*
6.10.5 Case Studies *372*
Selected Reading *382*
References *383*

6.11 Chemoinformatics at the CADD Group of the National Cancer Institute *385*
Megan L. Peach and Marc C. Nicklaus
6.11.1 Introduction and History *385*
6.11.2 Chemical Information Services *386*
6.11.3 Tools and Software *388*
6.11.4 Synthesis and Activity Predictions *391*
6.11.5 Downloadable Datasets *391*
References *392*

6.12 Uncommon Data Sources for QSAR Modeling *395*
Alexander Tropsha
6.12.1 Introduction *395*
6.12.2 Observational Metadata and QSAR Modeling *397*
6.12.3 Pharmacovigilance and QSAR *398*
6.12.4 Conclusions *401*
Selected Reading *402*
References *402*

6.13 Future Perspectives of Computational Drug Design *405*
Gisbert Schneider
6.13.1 Where Do the Medicines of the Future Come from? *405*
6.13.2 Integrating Design, Synthesis, and Testing *408*
6.13.3 Toward Precision Medicine *409*
6.13.4 Learning from Nature: From Complex Templates to Simple Designs *411*
6.13.5 Conclusions *413*
Selected Reading *414*
References *414*

7 Computational Approaches in Agricultural Research *417*
Klaus-Jürgen Schleifer
7.1 Introduction *417*
7.2 Research Strategies *418*
7.2.1 Ligand-Based Approaches *419*
7.2.2 Structure-Based Approaches *422*
7.3 Estimation of Adverse Effects *429*
7.3.1 *In Silico* Toxicology *429*
7.3.2 Programs and Databases *430*
7.3.3 *In Silico* Toxicology Models *432*
7.4 Conclusion *435*
Selected Reading *436*
References *436*

8 Chemoinformatics in Modern Regulatory Science *439*
Chihae Yang, James F. Rathman, Aleksey Tarkhov, Oliver Sacher, Thomas Kleinoeder, Jie Liu, Thomas Magdziarz, Aleksandra Mostraq, Joerg Marusczyk, Darshan Mehta, Christof Schwab, and Bruno Bienfait
8.1 Introduction *439*
8.1.1 Science and Technology Progress *439*
8.1.2 Regulatory Science in Twenty-First Century *440*
8.2 Data Gap Filling Methods in Risk Assessment *441*
8.2.1 QSAR and Structural Knowledge *442*
8.2.2 Threshold of Toxicological Concern (TTC) *443*
8.2.3 Read-Across (RA) *445*
8.3 Database and Knowledge Base *448*
8.3.1 Architecture of Structure-Searchable Toxicity Database *448*

8.3.2 Data Model for Chemistry-Centered Toxicity Database *449*
8.3.3 Inventories *452*
8.4 New Approach Descriptors *453*
8.4.1 ToxPrint Chemotypes *453*
8.4.2 Liver BioPath Chemotypes *458*
8.4.3 Dynamic Generation of Annotated Linear Paths *459*
8.4.4 Other Examples of Descriptors *461*
8.5 Chemical Space Analysis *462*
8.5.1 Principal Component Analysis *462*
8.6 Summary *464*
Selected Reading *466*
References *466*

9 Chemometrics in Analytical Chemistry *471*
Anita Rácz, Dávid Bajusz, and Károly Héberger
9.1 Introduction *471*
9.2 Sources of Data: Data Preprocessing *472*
9.3 Data Analysis Methods *475*
9.3.1 Qualitative Methods *475*
9.3.2 Quantitative Methods *483*
9.4 Validation *488*
9.5 Applications *492*
9.6 Outlook and Prospects *492*
Selected Reading *496*
References *496*

10 Chemoinformatics in Food Science *501*
Andrea Peña-Castillo, Oscar Méndez-Lucio, John R. Owen, Karina Martínez-Mayorga, and José L. Medina-Franco
10.1 Introduction *501*
10.2 Scope of Chemoinformatics in Food Chemistry *502*
10.3 Molecular Databases of Food Chemicals *503*
10.4 Chemical Space of Food Chemicals *506*
10.4.1 General Considerations *506*
10.4.2 Chemical Space Analysis of Food Chemical Databases *508*
10.5 Structure–Property Relationships *510*
10.5.1 Structure–Flavor Relationships and Flavor Cliffs *511*
10.5.2 Quantitative Structure–Odor Relationships *512*
10.6 Computational Screening and Data Mining of Food Chemicals Libraries *513*
10.6.1 Anticonvulsant Effect of Sweeteners and Pharmaceutical and Food Preservatives *514*
10.6.2 Mining Food Chemicals as Potential Epigenetic Modulators *516*
10.7 Conclusion *521*
Selected Reading *522*
References *523*

11 Computational Approaches to Cosmetics Products Discovery *527*
Soheila Anzali, Frank Pflücker, Lilia Heider, and Alfred Jonczyk
11.1 Introduction: Cosmetics Demands on Computational Approaches *527*
11.2 Case I: The Multifunctional Role of Ectoine as a Natural Cell Protectant (Product: Ectoine, "Cell Protection Factor", and Moisturizer) *528*
11.2.1 Molecular Dynamics (MD) Simulations *530*
11.2.2 Results and Discussion: Ectoine Retains the Power of Water *531*
11.3 Case II: A Smart Cyclopeptide Mimics the RGD Containing Cell Adhesion Proteins at the Right Site (Product: Cyclopeptide-5: Antiaging) *533*
11.3.1 Methods *536*
11.3.2 Results and Discussion *536*
11.4 Conclusions: Cases I and II *542*
References *545*

12 Applications in Materials Science *547*
Tu C. Le, and David A. Winkler
12.1 Introduction *547*
12.2 Why Materials Are Harder to Model than Molecules *548*
12.3 Why Are Chemoinformatics Methods Important Now? *548*
12.4 How Do You Describe Materials Mathematically? *549*
12.5 How Well do Chemoinformatics Methods Work on Materials? *551*
12.6 What Are the Pitfalls when Modeling Materials? *551*
12.7 How Do You Make Good Models and Avoid the Pitfalls? *553*
12.8 Materials Examples *554*
12.8.1 Inorganic Materials and Nanomaterials *554*
12.8.2 Polymers *557*
12.8.3 Catalysts *558*
12.8.4 Metal–Organic Frameworks (MOFs) *560*
12.9 Biomaterials Examples *561*
12.9.1 Bioactive Polymers *561*
12.9.2 Microarrays *564*
12.10 Perspectives *566*
Selected Reading *567*
References *567*

13 Process Control and Soft Sensors *571*
Kimito Funatsu
13.1 Introduction *571*
13.2 Roles of Soft Sensors *573*
13.3 Problems with Soft Sensors *574*
13.4 Adaptive Soft Sensors *576*

13.5 Database Monitoring for Soft Sensors *578*
13.6 Efficient Process Control Using Soft Sensors *581*
13.7 Conclusions *582*
Selected Readings *583*
References *583*

14 Future Directions *585*
Johann Gasteiger
14.1 Well-Established Fields of Application *585*
14.2 Emerging Fields of Application *586*
14.3 Renaissance of Some Fields *587*
14.4 Combined Use of Chemoinformatics Methods *588*
14.5 Impact on Chemical Research *589*

Index *591*