

# Contents

Preface — v

List of figures — xiv

List of tables — xviii

List of symbols — xxi

**Introduction and outline — 1**

<b>1</b>	<b>Basics of graphs and molecular graphs — 13</b>
1.1	Graphs — 13
1.1.1	Labeled graphs — 14
1.1.2	Unlabeled graphs — 17
1.2	Molecular graphs, constitutional isomers — 25
1.2.1	Atom states in organic chemistry — 26
1.2.2	Constitutional isomers — 29
1.2.3	The existence of molecular graphs — 34
1.3	Group actions on molecular graphs — 35
1.3.1	Counting unlabeled structures — 36
1.3.2	Counting by weight — 44
1.3.3	Constructive methods — 47
1.3.4	Generating samples — 52
<b>2</b>	<b>Advanced properties of molecular graphs — 56</b>
2.1	Substructures — 56
2.1.1	Graph-theoretical elements — 56
2.1.2	Subgraphs and their embeddings — 59
2.2	Molecular substructures — 62
2.2.1	Ambiguous molecular graphs — 62
2.2.2	Substructure restrictions — 64
2.3	Chemical reactions — 66
2.4	Mesomerism — 69
2.5	Existing chemical compounds — 72
2.6	Molecular descriptors — 76
2.6.1	Arithmetical descriptors — 78
2.6.2	Topological descriptors — 79
2.6.3	Geometrical descriptors — 87

**3 Chirality — 91**

- 3.1 Orientation and chirality — 92
  - 3.1.1 Barycentric placement of molecules in space — 93
  - 3.1.2 Symmetry operations, the point group — 98
  - 3.1.3 Chirality and handedness — 102
- 3.2 Permutational isomers — 106
  - 3.2.1 Counting permutational isomers — 109
  - 3.2.2 Permutational isomers by content — 113
  - 3.2.3 Enumeration by symmetry — 118
  - 3.2.4 Constructive aspects — 124

**4 Stereoisomers — 132**

- 4.1 Basic stereochemistry — 132
  - 4.1.1 Symmetry, the orientational automorphism group — 140
  - 4.1.2 Partial orientation functions (POFs) — 141
  - 4.1.3 Generation of abstract POFs — 143
  - 4.1.4 Tests for chemical realizability — 146
- 4.2 Radon partitions — 150
- 4.3 Binary Grassmann–Plücker relations — 154
- 4.4 Chemical conformation and cyclohexane — 158
- 4.5 Perspectives — 162

**5 Molecular structure generation — 164**

- 5.1 Formula-based structure generation — 165
  - 5.1.1 Orderly generation of simple graphs — 165
  - 5.1.2 Introducing constraints — 171
  - 5.1.3 Variations and refinements — 172
  - 5.1.4 From simple graphs to multigraphs — 173
  - 5.1.5 Applying the Homomorphism Principle — 174
  - 5.1.6 Orderly generation — 176
  - 5.1.7 Beyond orderly generation — 179
- 5.2 Constrained generation and fuzzy formulas — 180
  - 5.2.1 Restrictions for a molecular formula — 181
  - 5.2.2 Structural restrictions — 182
- 5.3 Reaction-based structure generation — 183
  - 5.3.1 Libraries of permutational isomers — 183
  - 5.3.2 Attaching substituents to a central molecule — 190
  - 5.3.3 Generation using the network principle — 191
  - 5.3.4 Generation of MS fragments — 193
  - 5.3.5 Construction using the network principle — 194
  - 5.3.6 Combinatorial libraries — 195
  - 5.3.7 Ugi's seven component reaction — 196

- 5.4 Generic structural formulas — 199
  - 5.4.1 A simple generic structural formula — 199
  - 5.4.2 Patents in chemistry — 202
- 5.5 Canonizing molecular graphs — 204
  - 5.5.1 Initial classification — 206
  - 5.5.2 Iterative refinement — 207
  - 5.5.3 Labeling by backtracking — 209
  - 5.5.4 Pruning the backtrack tree — 210
  - 5.5.5 Profiting from symmetry — 214
- 5.6 Data structures for molecular graphs — 219
  
- 6 Supervised statistical learning — 221**
  - 6.1 Variables and predicting functions — 221
    - 6.1.1 Regression and classification — 222
    - 6.1.2 Validation of the predicting function — 224
    - 6.1.3 Preprocessing of data — 227
    - 6.1.4 Selection of variables — 228
  - 6.2 Models for predicting functions — 231
    - 6.2.1 Linear models — 231
    - 6.2.2 Neural networks — 233
    - 6.2.3 Support vector machines — 234
    - 6.2.4 Decision trees — 236
    - 6.2.5 Nearest neighbors — 238
  
- 7 Quantitative structure–property relationships — 240**
  - 7.1 Optimization of experiments in combinatorial chemistry — 240
  - 7.2 The use of molecular descriptors — 242
    - 7.2.1 Arithmetical, topological, and geometrical descriptors — 243
    - 7.2.2 Substructure counts — 250
  - 7.3 Mathematical composition of QSPRs — 251
  - 7.4 Case studies of QSPRs obtained by linear modeling — 254
    - 7.4.1 Linear modeling using topological indices — 255
    - 7.4.2 Linear modeling using substructure counts — 261
    - 7.4.3 Linear modeling using TI and SC — 265
    - 7.4.4 Further descriptors and regression methods — 268
    - 7.4.5 Prediction — 270
  - 7.5 Case studies with separate learning and test sets — 270
    - 7.5.1 Preprocessing of structures — 271
    - 7.5.2 Choice of descriptors — 273
    - 7.5.3 Linear modeling by best subset selection — 275
    - 7.5.4 Linear modeling by stepwise subset selection — 277
    - 7.5.5 Linear modeling using principal component regression — 282

## List of figures

- Fig. 1 Seven pairs of  $C_6H_6$  stereoisomers. — 4
- Fig. 2 Two pairs of  $C_6H_6$  conformers. — 5
- Fig. 2.1 Constitutional isomers of  $C_6H_6$  in *Beilstein*. — 73
- Fig. 2.2 Steric energy of the constitutional isomers  $C_6H_6$ . — 74
- Fig. 2.3 Van der Waals volumes of the constitutional isomers  $C_6H_6$ . — 75
- Fig. 3.1 The 22 permutational isomers of Seveso dioxin. — 130
- Fig. 4.1 Three chiral compounds whose stereoisomerism cannot be described in terms of stereocenters, stereogenic double bonds, or rotatable single bonds. — 133
- Fig. 4.2 Example conformations. — 139
- Fig. 4.3 Stereoisomers generated for structure 4.2a. — 145
- Fig. 4.4 Conformers generated for structure 4.2b. — 145
- Fig. 4.5 Kinds of quadruples considered in tests. — 146
- Fig. 4.6 Stereoisomers and conformers generated for structure 4.2c. — 148
- Fig. 4.7 Radon partitions. — 151
- Fig. 4.8 Two chemically forbidden minimal Radon partitions. — 157
- Fig. 4.9 A non-chemical arrangement of atoms for structure b, Figure 4.2. — 158
- Fig. 4.10 The chair form of cyclohexane with numbered atoms. — 158
- Fig. 4.11 Two forbidden minimal Radon partitions. — 160
- Fig. 4.12 Four conformations found by constrained optimization. — 160
- Fig. 5.1 Backtrack tree for labeled generation of simple graphs on three nodes. — 169
- Fig. 5.2 Backtrack tree for unlabeled generation of simple graphs on three nodes. — 170
- Fig. 5.3 Backtrack tree for orderly generation of simple graphs on three nodes. — 171
- Fig. 5.4 Adjacency matrix with block structure. — 177
- Fig. 5.5 Structures of the 20 proteinogenic amino acids. — 186
- Fig. 5.6 Scheme of the seven component reaction. — 198
- Fig. 5.7 Alkyl groups with 1–6 C atoms. — 202
- Fig. 5.8 Compounds 1–4 with arbitrary initial vertex numbering. — 207
- Fig. 5.9 The backtrack tree for structure 3, Figure 5.8. — 214
- Fig. 5.10 The backtrack tree for cubane. — 216